

Science's new social contract with society

Michael Gibbons

Under the prevailing contract between science and society, science has been expected to produce 'reliable' knowledge, provided merely that it communicates its discoveries to society. A new contract must now ensure that scientific knowledge is 'socially robust', and that its production is seen by society to be both transparent and participative.

Modern science has until recently flourished partly because of a stable, underlying agreement between its practitioners and the rest of society. In other words, there has been a social contract between science and society, an arrangement built on trust which sets out the expectations of the one held by the other, and which — in principle — includes appropriate sanctions if these expectations are not met.

This social contract has been made up of several individual elements, reflecting broader contracts between government and society, between industry and society, and between higher education and society. The contract between university science and society, for example, has been based traditionally on the understanding that universities will provide research and teaching in return for public funding and a relatively high degree of institutional autonomy; under this contract, the universities, often supported through research-funding agencies, have been expected to generate fundamental knowledge for society, and to train the highly qualified manpower required by an advanced industrial society.

By contrast, the contract with industrial research and development (R&D) has been based on an understanding that industry would provide for the appliance of science through the work of its laboratories, and thus carry the discoveries of basic science into product and process innovations. In turn, government science was meant to use research establishments to fill the gap between university science and industrial R&D. The understanding has been that the state has been directly responsible for carrying out research related to national need; for example, in defence, energy, public health and standards.

For most of the twentieth century, universities, government research establishments and industrial laboratories have therefore operated relatively independently, developing their own research practices and modes of behaviour. Recently, however, this relative institutional impermeability has gradually become more porous. Privatization policies,



Traditional boundaries between university and industrial science, and between basic and applied research, are disappearing. As a result, science and society are invading each other's domain, requiring a rethinking of previous responsibilities.

for example, have moved many government research establishments into the market place. With the relaxation of the Cold War, governments have shifted their priorities from security and military objectives to maintaining international competitiveness and enhancing the quality of life. And many long-established industries have been denationalized, while in many countries companies previously dependent upon government for R&D support through military technology projects have had to find these resources elsewhere, or in partnership with others, to compete in international markets.

Meanwhile the expansion of higher education has been accompanied by a culture of accountability that has impacted on both teaching and research. In research, many academics have had to accept objective-driven research programmes, whereas research funding agencies have been increasingly transformed from primarily responsive institutions, responsible for maintaining basic science in the universities, into instruments for attaining national technological, economic and social priorities through the funding of research projects and programmes.

These trends can be observed internationally, even if their precise form and timing has varied between countries. Cumulatively, they signal the end of the institutional arrangements through which science flourished during and after the Second World War, and thus mark the expiry of the social contract between science and society that has dominated this period. A new social contract is now required. This cannot be achieved merely by patching up the existing framework. A fresh approach — virtually a complete 'rethinking' of science's relationship with the rest of society — is needed.

Reflecting complexity and diversity

One aspect of this new contract is that it needs to reflect the increasing complexity of modern society. For example, there are no longer clear demarcation lines between university science and industrial science, between basic research, applied research and product development, or even between careers in the academic world and in industry. There is now greater movement across institutional boundaries, a blurring of professional identities and a greater diversity of career patterns.

But the price of this increased complexity is a pervasive uncertainty. One way of looking at this is in terms of an erosion of society's stable categorizations, namely the state, market, culture and science. Alternatively, it can be seen as the cumulative effect of parallel evolutionary processes. For there has been a co-evolution in both society and science in terms of the range of organizations with which researchers are prepared to work, the colleagues with whom they collaborate, and topics considered interesting. Whatever viewpoint one takes, science is now produced in more open systems of knowledge production.

One consequence is that the norms and practices of research in university and industrial laboratories have converged. There are still differences between universities and industry, but these do not impact on what is considered sound scientific practice¹. Indeed, science and society more generally





Both pressure groups and ordinary consumers are demanding that the debate surrounding the health implications of GMOs be broadened to include the perspectives of the non-expert community.

have each invaded the other's domain, and the lines demarcating the one from the other have virtually disappeared.

As a result, not only can science speak to society, as it has done so successfully over the past two centuries, but society can now 'speak back' to science. The current contract between science and society was not only premised on a degree of separation between the two, but also assumed that the most important communication was from science to society. Science was seen as the fountainhead of all new knowledge and, as part of the contract, was expected to communicate its discoveries to society. Society in turn did what it could to absorb the message and through other institutions — primarily industry — to transform the results of science into new products and processes.

Science was highly successful working in this mode, and for as long as it delivered the goods, its autonomy was seldom contested. Yet this success has ironically itself been instrumental in changing its relationship with society, drawing science into a larger and a more diverse range of problem areas, many lying outside traditional disciplinary boundaries. It is this increasingly intense involvement of science in society over the past half a century that has created the conditions that underpin the growing complexity and the pervasive uncertainty in which we live, and encouraged the social and behavioural experiments described above.

But if it is widely recognized that science is transforming modern society, it is less often appreciated that society, in speaking back, is transforming science. I will use the term 'contextualization' to describe this process, and 'contextualized knowledge' as the outcome of this reverse communication. Contextualization affects modern science in

its organization, division of labour and day-to-day practices, and also in its epistemological core.

In relation to the former, for example, research carried out in both industrial and government laboratories, as well as the funding policies of research-funding agencies, have opened up to a wide range of socioeconomic demands, admitting more and more cross-institutional links, and thus altering the balance between the different sources of funding of academic research. Thus in 'speaking back' to science, society is demanding various innovations, for example the pursuit of national objectives, the contribution to new regulatory regimes and acknowledgement of the multiplication of user-producer interfaces.

In relation to the latter, the epistemological dimension, the increasing importance of 'context' is also reflected in a relatively rapid shift within science from the search for 'truth' to the more pragmatic aim of providing a provisional understanding of the empirical world that 'works'². John Ziman, former physicist and long-time contributor to social studies of science, has described science as a form of 'reliable knowledge' that becomes established not in terms of an abstract notion of objectivity but, concretely, in terms of the replicability of research statements and the formation of a consensus within the relevant peer group³. Reliable knowledge is therefore defined as such because it 'works'.

But what 'works' has now acquired a further dimension that can best be described as a shift from 'reliable knowledge' to what Nowotny *et al.* call 'socially robust knowledge'⁴. The latter characterization is intended to embrace the process of contextualization. For 'socially robust' knowledge has three

aspects. First, it is valid not only inside but also outside the laboratory. Second, this validity is achieved through involving an extended group of experts, including lay 'experts'. And third, because 'society' has participated in its genesis, such knowledge is less likely to be contested than that which is merely 'reliable'.

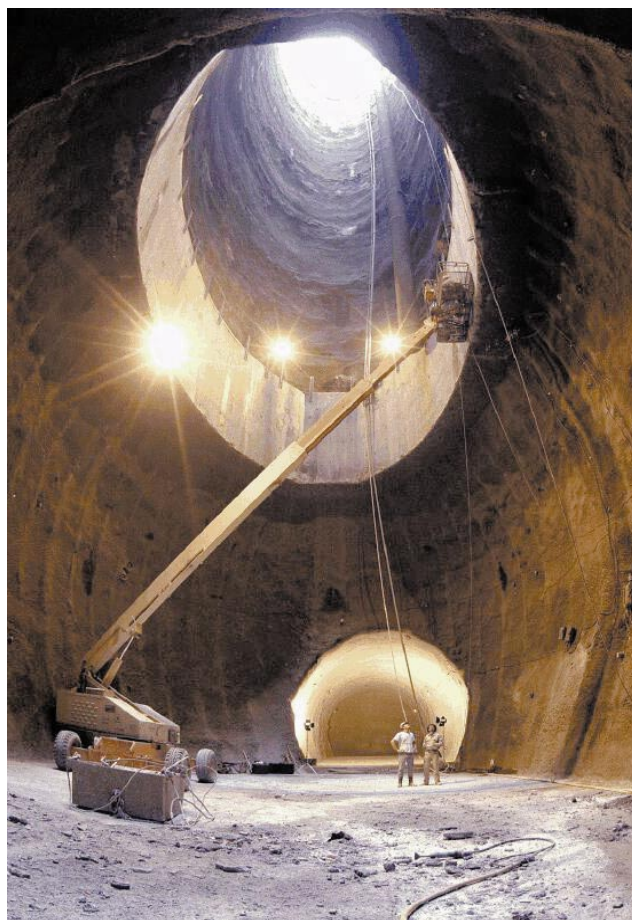
Socially robust knowledge

My argument is that we are currently witnessing a significant shift from 'reliable' to 'socially robust' knowledge. Three observations can immediately be made. The first is that the basic conditions and processes that have underpinned the production of 'reliable knowledge' are not necessarily compromised by the shift to 'socially robust knowledge'. Indeed, if these conditions and processes have been undermined, it may have been as much by the narrow outlook of much scientific practice as by any attempt to widen the range of stakeholders, or more systematically to take into account the context in which science is produced.

The second observation is that reliable knowledge has always only been reliable within boundaries. Science was recognized as inherently incomplete because it is, primarily, a method rather than a final answer. But to achieve a reasonable degree of reliability, the problem terrain had also to be circumscribed, and judgements on what is included there restricted to those of a peer group, rather than opened to the scientific community as a whole.

Both aspects of reliable knowledge are carried forward into socially robust knowledge. But although knowledge remains incomplete, this is no longer only in the conventional sense that it will eventually be superseded by superior science; rather it means that it may be sharply contested, and no longer remains within the controlled environment of scientific peers. This shift involves renegotiating and reinterpreting boundaries that have been dramatically extended, so that science can no longer not be validated as reliable by conventional discipline-bound norms; while remaining robust, science must now be sensitive to a much wider range of social implications.

An example is the current debate surrounding genetically modified organisms (GMOs). Here, specialist peer groups have been challenged not only by pressure groups but also by ordinary consumers, for whom the research process is far from transparent, and who are demanding that it be more so. Here, knowledge of the health implications of GMOs may be 'reliable' in the conventional scientific sense; but it is not socially robust, and will not become so until the peer group is broadened to take into account the perspectives and concerns of a much wider section of the community.



A failure to persuade the broad public of the value of the US Superconducting Super Collider research programme may have contributed to the collapse of funding for the project.

fragmentation. Such narratives are challenging to their participants. Experts must respond to issues and questions that are never merely scientific and technical, and must address audiences that never consist only of other experts. The limits of competence of the individual expert call for the involvement of a wide base of expertise that has to be carefully orchestrated if it is to speak in unison.

Since expertise now has to bring together knowledge that is itself distributed, contextualized and heterogeneous, it cannot arise at one specific site, or out of the views of one scientific discipline or group of highly respected researchers. Rather it must emerge from bringing together the many different 'knowledge dimensions' involved. Its authority depends on the way in which such a collective group is linked, often in a self-organized way. Breakdowns in social authority arise when links are inadequately established, as has occurred in European debates over GMOs.

Rethinking science

These four inter-related processes — co-evolution, contextualization, the production of socially robust knowledge and the construction of narratives of expertise — form a framework both for rethinking science and for understanding any new social contract between society and science. Co-evolution denotes an open interaction between science and society which generates variety through experimentation, whether in scientific problems, colleagues or institutional designs, with the selective retention of certain choices, modes or solutions. This is so even while these experimental approaches, in responding to uncertainty and complexity, not only promote permeability but also generate more complexity and uncertainty, thus encouraging further experimentation.

Greater permeability provides the basis for increased contextualization by increasing the routes through which society can 'speak back' to science. Denser communication itself brings an imperative to produce socially robust knowledge that is seen as valid not only inside but also outside the walls of the laboratory, in terms of being accepted as legitimate.

As the walls of laboratories have opened up, more and more scientists have taken their places as actors in the agora, broadening the range of experts whose view might be sought on a particular problem or issue. To cope with this, a further development in the use of scientific and technical experts is needed. For reliable knowledge can only become socially robust if society sees the process of knowledge production as transparent and participative. The old image of science working autonomously will no longer suffice. Rather, a reciprocity is required in which not only does the public understand how science

There was also a degree of contestation in the United States about the value of the Superconducting Super Collider (SSC), plans for which were dropped in 1992. In this case, however, unlike the case with GMOs, there was no spontaneous backlash from society generally about the value of the knowledge. Rather, it has been argued that the collapse of funding for the project was a result of the unwillingness (or inability) of a narrow disciplinary group to extend its boundaries sufficiently to persuade other scientists and politicians that the research would be of wide benefit⁵. Again we see a failure to achieve sufficient social robustness in the research process, however reliable it may be in its own terms.

The third observation is that the epistemological core of science has, over time, become crowded with norms and practices that cannot be reduced easily to a single generic methodology, or, more broadly, to privileged cultures of scientific inquiry. There is no one set of practices that describe, much less lead to, good science. The case for science can still be made in essentially functionalist terms; but many more factors now need to be taken into account before a solution that 'works' can be adopted.

One outcome of all these changes is that the sites at which problems are formulated and negotiated have moved from their previous institutional locations in government,

industry and universities into the 'agora' — the public space in which both 'science meets the public' and the public 'speaks back' to science. This is a space in which the media is increasingly active, and in which the new communication technologies play a prominent role. It is also the domain in which contextualization occurs. Neither state nor market, neither exclusively private nor exclusively public, the agora is where today's societal and scientific problems are framed and defined, and their 'solutions' are negotiated.

Narratives of expertise

The factor that has come to the fore in the agora is the role of scientific and technical expertise that is so crucial to decision making in highly industrialized societies. This role is changing as expertise spreads throughout society, resulting in the fragmentation of established links between expertise and institutional structures, whether of government, industry or the professions. Furthermore, the questions asked of experts are neither the same, nor simple extensions of, the ones that arise in their specialist fields of study. Experts must now extend their knowledge to widely disparate areas, and try to integrate what they 'know' now with what others want to 'do' in the future.

Collective narratives of expertise need to be constructed to deal with the complexity and the uncertainty generated by this

works but, equally, science understands how its publics work.

The process of rethinking science, now that the line that used to separate science and society has virtually disappeared, has scarcely begun⁵. But several changes in perspective must be initiated before a new social contract can emerge. First, the need for contextualization means that the (unknowable) implications as well as the (planned or predictable) applications of scientific research have to be embraced.

Research activities now transcend the immediate context of application, and begin to reach out, anticipate and engage reflexively with those further entanglements, consequences and impacts that it generates. This 'context of implication' always transcends the immediate 'context of application' in which it occurs. It may embrace neighbouring research fields, and as yet obscurely recognized future uses. Taking the 'context of implication' seriously opens the door to those previously excluded from decisions about research. The individuals now involved may be encountered haphazardly as individuals, perhaps colleagues or rivals. They may come from other scientific disciplines, or from the 'user' community. Whatever their origins, scientific knowledge will increasingly need to be tested not only against nature, but against (and hopefully also with) other people.

Furthermore, while it is important to define problems, and then assemble the intellectual, human and financial resources needed to solve them, this is not, in itself, sufficient to guarantee the reflexivity characteristic of socially distributed knowledge production. In contrast, a process of contextualization that attempts to embrace unpredictable and unintended implications demands reflexivity, as it is intended to incorporate future potential implications into the research process from the very beginning. It thus goes far beyond a conventional 'forward look' or 'technology foresight' exercise.

This has several consequences. One is the need for strategies to 'fix' more accurately the implications of knowledge production. This might be done by identifying areas in which significant implications of particular research projects are likely to arise without being pinpointed exactly, making it necessary to 'prospect' for these (presently unknowable) implications. Such a process might, for example, involve consulting other knowledge producers and users, as well as wider social constituencies, in order to carry out a form of 'triangulation' survey. Perhaps every research proposal and project should include a deliberate strategy for identifying its 'context of implication'. This might best be achieved by including those likely to be implicated — perhaps unknowingly — as well as the conscious carriers of social knowledge.

A second need is for the process of contextualization to be internalized. The 'context of application' can be managed through 'external' mechanisms such as 'forward look' and 'technology foresight' exercises, and through science and technology parks and technology transfer or industrial liaison offices within universities. In contrast, the 'context of implication' needs to be internalized by researchers if it is to be effective. It is expressed through routes, often informal, that cannot easily be incorporated into administrative procedures. These new lines of communication need to be encouraged and recognized institutionally. This cannot be done by communications experts — contextualization is not a public relations exercise — or by asking journalists to develop popular accounts of the significance of research.

A further important point is that the more open and comprehensive the scientific community, the more socially robust will be the knowledge it produces. This is contrary to the traditional assumption that there is a strong relationship between the social and intellectual coherence (and, therefore, the boundedness) of a scientific community, and the reliability of the knowledge it produces. Reliable knowledge may have been best produced by such cohesive (and therefore restricted) scientific communities. But socially robust knowledge can only be produced by much more sprawling socio/scientific constituencies with open frontiers.

At the same time, socially robust knowledge is superior to reliable knowledge both because it has been subject to more intensive testing and retesting in many more contexts — which is why it is 'robust' — and also because of its malleability and connective capability. Its context is not predetermined or fixed, but open to ceaseless renegotiation. Instead of achieving a precarious invariance by establishing strict limits within which its truthfulness can be tested, as reliable knowledge does, socially robust knowledge is the product of an intensive (and continuous) interaction between data and other results, between people and environments, between applications and implications.

It is also clear that science must leave the ivory tower and enter the agora. To increase the effectiveness with which the agora operates, the self-organizing capacity of all participants needs to be enhanced. Here, there is tension between the desire for individual or institutional autonomy and the increasing demands for accountability on both individuals and institutions.

Increasing the capacity for self-organizing means that participants need to act more reflexively. But one cannot enhance responsiveness by simply increasing the demand for public accountability, as this could make participants more, rather than less, defensive. On the contrary, what is needed is to encourage participants voluntarily to

internalize accountability. Indeed there is an analogy between the relationships between autonomy and self-organization on the one hand, and between reliable and socially robust knowledge on the other. In the agora, the conditions that promote greater self-organization also promote the generation of socially robust knowledge.

Conclusion

To summarize, I have argued in this paper that the prevailing contract between science and society was set up to sustain the production of 'reliable knowledge'; a new one must ensure the production of 'socially robust knowledge'. The prevailing contract is governed by the rules of bureaucratic rationality, with society linked to 'people' primarily through representative institutions. A new contract will require more open, socially distributed, self-organizing systems of knowledge production that generate their own accountability and audit systems. Under the prevailing contract, science was left to make discoveries and then make them available to society. A new contract will be based upon the joint production of knowledge by society and science.

A new social contract will therefore involve a dynamic process in which the authority of science will need to be legitimated again and again. To maintain this, science must enter the agora and participate fully in the production of socially robust knowledge. According to some observers, we can already see this approach emerging in the management of large technology projects. Thomas P. Hughes, the eminent American historian of technology, has identified a new ethos among engineers who now recognize that the deeper involvement of communities in decision making actually produced better engineering solutions in a number of projects⁶. If the boundaries between science, technology and society are becoming more permeable, why should not a similar approach in science likewise produce more socially robust solutions?

Michael Gibbons, a former director of the Science Policy Research Unit at the University of Sussex, is now secretary-general of the Association of Commonwealth Universities, 36 Gordon Square, London WC1H 0PF, UK.

1. van Duinen, R. J. European research councils and the Triple Helix. *Sci. Public Policy* 25, 381–386 (1998).
2. Daston, L. & Galison, P. The image of objectivity. *Representations* 40, 81–128 (1992).
3. Ziman, J. *Reliable Knowledge* Canto Edn (Cambridge Univ. Press, Cambridge, 1991).
4. Nowotny, H., Scott, P. & Gibbons, M. *Re-thinking Science: Knowledge Production in an Age of Uncertainty* (in the press).
5. Gieryn, T. F. in *Handbook of Science and Technology Studies* (eds Jasanoff, S. et al.) 393–443 (Sage Publishers, London, 1995).
6. Hughes, T. P. *Rescuing Prometheus* 301–303 (Pantheon Books, New York, 1998).

Acknowledgements. The research described in this essay is taken from a report entitled 'Re-thinking science: knowledge production in a Mode 2 Society', prepared by H. Nowotny, P. Scott and M. Gibbons for, and funded by, the Tercentenary Fund of the Royal Swedish Bank and the Swedish Council for Higher Education.

How common are habitable planets?

Jack J. Lissauer

The Earth is teeming with life, which occupies a diverse array of environments; other bodies in our Solar System offer fewer, if any, niches that are habitable by life as we know it. Nonetheless, astronomical studies suggest that many habitable planets may be present within our Galaxy.

One of the most basic questions that has been pondered by Natural Philosophers for many millennia concerns humanity's place in the Universe: are we alone? This question has been approached from many different viewpoints, and similar reasoning has led to widely diverse answers. Copernicus, Kepler, Galileo and Newton each demonstrated convincingly that other planets that were qualitatively similar to Earth orbit the Sun. In the past few years, more than 20 planets have been discovered in orbit about stars other than our Sun; these are 'extrasolar' planets.

The intellectual and technological advances of the past century leave us poised at the turn of the millennium to investigate the possibility of extraterrestrial life along numerous paths of experimental, observational and theoretical studies in both the physical and life sciences.

Prerequisites for habitability

Here I assume that extraterrestrial life would be carbon-based and use liquid water (characteristics common to all life found on Earth), and I define a 'habitable planet' as one capable of supporting such life.

Life on Earth has been able to evolve and thrive thanks to billions of years of benign climate. Mars seems to have had a climate sufficiently mild for liquid water to have flowed on its surface when the Solar System was roughly one-tenth its current age, but at present, its low atmospheric pressure means that liquid water is not stable on the martian surface. Venus is too hot, with a massive atmosphere dominated by carbon dioxide; we cannot say whether or not young Venus had a mild Earth-like climate. Indeed, as models of stellar evolution predict that the young Sun was about 25 per cent less luminous than at present, we do not understand why Earth, much less Mars, was warm enough to be covered by liquid oceans 4 billion years ago, when life is thought to have originated.

Carbon dioxide is important for carbon-based life. On Earth, this compound cycles

Since one of the most wondrous and noble questions in Nature is whether there is one world or many, a question that the human mind desires to understand, it seems desirable for us to inquire about it.

Albertus Magnus, 13th century

— on a wide range of timescales — between the atmosphere, the oceans, living organisms, fossil fuels and carbonate rocks. The carbonate rocks form the largest reservoir, and are produced by reactions involving water (and in some cases living organisms). Carbon dioxide is recycled from carbonates back into the atmosphere as tectonic plates descend into the Earth's mantle and are heated. Carbonates are not readily recycled on a geologically inactive planet such as Mars, and they are not formed on planets like Venus, which lack surface water. Larger planets of a given composition remain geologically active for longer, as they have smaller ratios of surface area to mass, and thus retain heat from accretion and radioactive decay for longer. The number of variables involved in determining a planet's habitability precludes a complete discussion, but some of the main issues are summarized in Fig. 1 (compare ref. 1).

Stellar properties and habitability

Stars are huge balls of plasma that radiate energy from their surfaces and liberate energy through thermonuclear fusion reactions in their interiors. During a star's long-lived

'main-sequence phase', hydrogen in its core is gradually 'burned up' to maintain sufficient pressure to balance gravity. The star's luminosity (energy output) grows slowly during this phase, as fusion increases the mean particle mass in the core and greater temperature is required to achieve pressure balance. Once the hydrogen in the core is used up, the star's structure and luminosity change much more rapidly. Both Sun-like stars and larger stars expand and become 'red giants'; those stars with an initial mass greater than about eight solar masses can end their lives in spectacular supernova explosions.

What stars are most likely to have habitable planets? To make the astronomical problem more straightforward, I assume that the single factor required for supporting life on a planet is the presence of water on its surface over a long timescale. The main-sequence phase of low-mass stars (such as our Sun) provides 'continuously habitable zones'; when in orbits within these zones, planets may maintain liquid water on their surfaces for billions of years. High-mass stars are much hotter than low-mass stars and use up their fuel far more rapidly. Thus, even if Earth-like planets form around high-mass stars at distances where liquid water is stable, it is unlikely that benign conditions exist for long enough on these planets to enable life to form and evolve. However, the greater flux of ultraviolet radiation may speed up biological evolution enough to compensate for the shorter lifetime of a moderately massive star. At the other end of the size spectrum, the smallest, faintest stars can live for trillions of years, but they emit almost all of their luminosity at infrared wavelengths and their luminosity varies by tens of per cent owing to flares and large 'starspots' (analogous to sunspots). In addition, habitable-zone planets orbit so close to these faint stars that their rotation is tidally synchronized (as the Moon's rotation is relative to Earth); thus no day-night cycle occurs, and if the planet's atmosphere is thin it would freeze on the perpetually dark, cold hemisphere².

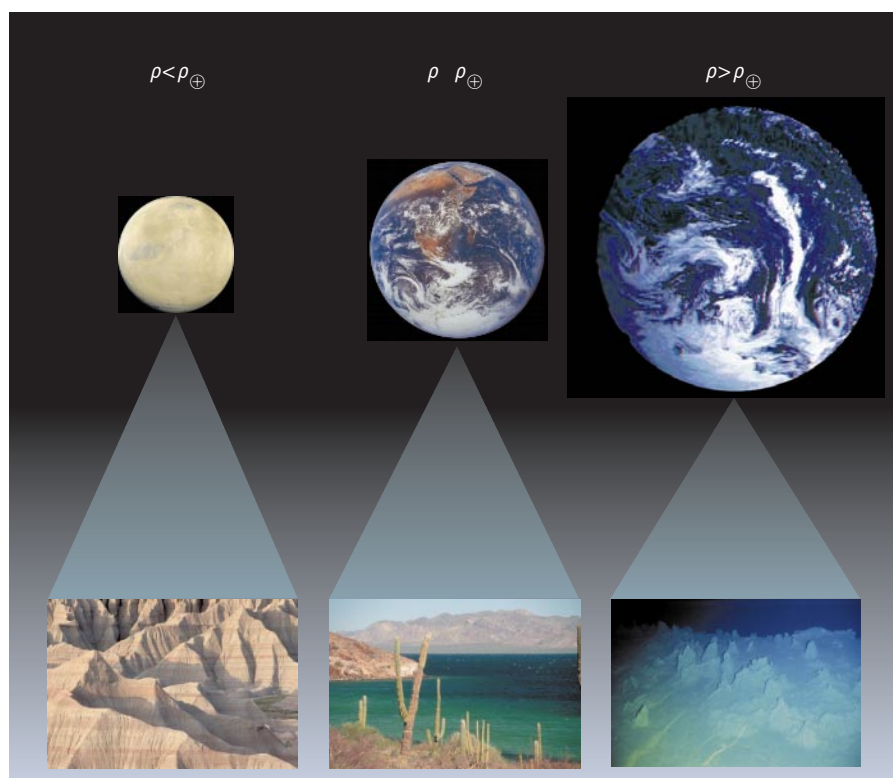


Figure 1 Environments on small and large Earth-like planets. Earth is depicted in the central panel of the figure. A smaller planet (left) made of the same material as Earth would be less dense, because the pressure in the interior would be lower. Such a planet would have a larger ratio of surface area to mass, so its interior would cool faster. Its lower surface gravity and more rigid crust would allow for higher mountains and deeper valleys than are seen on Earth. Most important to life is that it would have a much smaller surface pressure as a result of four factors: a larger ratio of surface area to mass, lower surface gravity, more volatiles sequestered in crust as there is less crustal recycling, and more atmospheric volatiles escaping to space. Among other things, this would imply a lower surface temperature, because there would be less greenhouse gases in the atmosphere. Some remedial measures that could improve the habitability of such a mass-deprived planet are: (1) move it closer to the star, so less greenhouse effect would be needed to keep the surface temperatures comfortable; (2) add extra atmospheric volatiles; and (3) include a larger fraction of long-lived radioactive nuclei than on Earth, to maintain crustal recycling. A larger planet (right) made of the same material as Earth would be denser and have a hotter interior. Its higher surface gravity and less rigid crust would lead to muted topography. It would have a much greater atmospheric pressure, and, unless its greenhouse effect was strong enough to boil away its water, much deeper oceans, probably covering its entire surface. Some remedial measures that could improve the habitability of such a mass-gifted planet are: (1) move it farther from the star; and (2) include a smaller fraction of atmospheric volatiles. It is not clear that more active crustal recycling would be a problem, within limits.

Stability of planetary systems

Although isolated single-planet systems are stable essentially indefinitely, mutual gravitational perturbations within multiple-planet systems can lead to collisions and ejections from the system. To a first approximation the star's gravity dominates, but planets exchange orbital energy and angular momentum, so that over millions or billions of orbits, even weak perturbations could move planets out of orbits that were initially in the habitable zone. Resonances among various orbital and precession (that is, rotation of the spatial orientation of the orbit ellipse) frequencies are the main source of chaos in planetary systems. There are no simple criteria for determining the stability of systems with many planets, but in general,

larger spacing between orbits, smaller eccentricities and inclinations, and lower-mass planets are more stable³.

One of these criteria — large spacing between orbits — has important implications. There is a minimum separation required for a system of Earth-mass planets to be stable for long periods of time. This separation is comparable to the width of a star's continuously habitable zone. Thus, arguments based on orbital stability support the possibility that most stars could have one or even two planets with liquid water on their surfaces. But unless greenhouse effects conspire to substantially compensate for increasing distance from the star, larger numbers of habitable planets per system are unlikely.

Planet formation

Stars are observed to be forming within cold regions of our Galaxy called molecular clouds. Even a very slowly rotating molecular-cloud core of stellar mass has far too much angular momentum to collapse down to an object of stellar dimensions. This leads to a (proto)star surrounded by a rotating disk of material; a significant fraction of the material in a collapsing core falls onto this disk. Such a disk has the same initial elemental composition as the growing star. At sufficient distances from the central star, it is cool enough for about 1–2 per cent of this material to be in solid form, either remnant interstellar grains or condensates formed within the disk. The growth from micrometre-sized dust to kilometre-sized solid bodies called planetesimals remains poorly understood⁴.

Kilometre-sized and larger planetesimals in protoplanetary disks travel on elliptical orbits that are altered by mutual gravitational interactions and physical collisions. These interactions lead to accretion (and in some cases, erosion and fragmentation) of planetesimals. Gravitational encounters within a 'swarm' of planetesimals can produce velocities that exceed the 'escape velocity' from the largest common planetesimals in the swarm⁵, and sufficiently massive and dense planets far enough from the star can hurl material into interstellar space. Comets in the Oort cloud — the vast comet reservoir ~5,000–50,000 AU from the Sun — are believed to be icy planetesimals that were sent outwards at nearly the Solar System escape velocity, and were perturbed into long-lived orbits around the Sun by close stars, interstellar clouds or the tidal forces of the Galactic disk.

We do not at present have enough data to determine the range of planetary systems that occur in nature. The gravitational pull of known extrasolar planets induce velocity variations in their stars much larger than would a planetary system like our own, and surveys so far accomplished have not detected low-mass and long-period planets⁶. Our own Solar System may represent a biased sample of a different kind, because it contains a planet with conditions suitable for life to evolve to the point of being able to ask questions about other planetary systems⁷. Unfortunately, we lack the capability to develop planet-formation models from first principles, or even from observations of protostellar disks, whose detailed properties are poorly known. But theory suggests that a great diversity of planetary systems are possible⁸.

We do, however, possess planet-formation models that can predict the growth times of giant planets (such as Jupiter). Current models⁹ predict growth times that are similar to estimates of the lifetime of gaseous protoplanetary disks. Thus, giant

Box 1 Interstellar eavesdropping

The Search for ExtraTerrestrial Intelligence (SETI) is an endeavour to detect signals from alien life forms²¹. A clear detection of such a signal would probably change humanity's world view as much as any other scientific discovery in history. As our society is in its technological infancy, another civilization capable of communicating over interstellar distances is likely to be enormously advanced compared with our own — compare our technology to that of a mere millennium ago and then extrapolate millions or billions of years into the future! Thus, a dialogue with extraterrestrials could alter our society in unimaginable ways (and they could probably answer most if not all of the questions raised in this article).

The primary instrument used by SETI is the radiotelescope. Most radio waves propagate with little loss through the interstellar medium, and many wavelengths also easily pass through the Earth's atmosphere. They are easy to generate and to detect. Radio thus appears to be an excellent means of interstellar communication, whether data are being exchanged between a community of civilizations around different stars or broadcast to the Galaxy in order to reach unknown societies in their technological infancy. Signals used for local purposes, such as radar and television on Earth, also escape and could be detected at great distances.

The first deliberate SETI radiotelescope observations were performed by Frank Drake in 1960. Since that time, improvements in receivers, data processing capabilities and radiotelescopes have doubled the capacity of SETI searches roughly every 8 months. Nonetheless, only a minuscule fraction of directions and frequencies have been searched, so one should not be discouraged by the lack of success so far.

planets might not form in most protoplanetary disks. Although giant planets themselves are unlikely abodes for life, they may harbour habitable moons. Moreover, they affect both the stability of the orbits of Earth-like planets and the flux of materials striking these planets⁷. Such impacts can have a devastating effect on life — a fact that no dinosaur is likely to argue with.

Impacts and dinosaurs

Impacts, like earthquakes, come in various sizes, with the large ones much rarer but

vastly more hazardous than the small ones (Table 1). Because of the destruction that impacts may produce, impact frequency is an important factor in planetary habitability. The impact rate on the terrestrial planets of our Solar System was orders of magnitude larger four billion years ago than it is at present. In a planetary system like our own, but with smaller planets replacing Jupiter and Saturn, large impact fluxes could continue, making planets with Earth-like compositions and radiation fluxes hostile abodes for living organisms⁷.

The largest mass extinction of the past 200 million years or so occurred 65 million years ago, when roughly half of the genera of multicellular organisms on Earth, including all of the dinosaurs, suddenly died off. The geological record shows a layer of impact-produced minerals and iridium, an element rare in the Earth's crust but more abundant in primitive meteorites, deposited at the time that the dinosaurs vanished (at the Cretaceous/Tertiary or K/T boundary). Additionally, the largest known crater on Earth dated at less than one billion years old was formed at this time. Taken together, these data imply that this K/T mass extinction was caused by the impact of an asteroid or comet, about 10 km in radius, into the Yucatan peninsula¹⁰.

A deluge of data to come

The speculation about advanced civilizations on Mars that was rampant a century ago has been quashed by better observations. However, evidence for a wet Mars billions of years ago suggests that life might have formed on that planet, and microbes may still be present below the surface¹¹. Interplanetary spacecraft will soon attempt to determine whether or not life ever existed on Mars, and if so what were (or are) its characteristics¹². NASA plans to investigate a possible subsurface ocean in Jupiter's moon Europa¹³, and the Cassini mission en route to Saturn will study the properties of Titan's atmosphere¹⁴. Although Titan's surface is too cold for liquid water, this large moon has a methane-rich atmosphere in which

Table 1 Impacts and life

Size	Example(s)	Most recent	Planetary effects	Effects on life
Super colossal $R > 2,000$ km	Moon-forming event	4.45×10^9 yr ago	Melts planet	Drives off volatiles; wipes out life on planet
Colossal $R > 700$ km	Pluto 1 Ceres (borderline)	$\approx 4.3 \times 10^9$ yr ago	Melts crust	Wipes out life on planet
Huge $R > 200$ km	4 Vesta (large asteroid)	$\sim 4.0 \times 10^9$ yr ago	Vaporizes oceans	Life may survive below surface
Extra large $R > 70$ km	Chiron (largest active comet)	3.8×10^9 yr ago	Vaporizes upper 100 m of oceans	Pressure-cooks photic zone; may wipe out photosynthesis
Large $R > 30$ km	Comet Hale-Bopp	$\sim 2 \times 10^9$ yr ago	Heats atmosphere and surface to $\sim 1,000$ K	Continents cauterized
Medium $R > 10$ km	K/T impactor 433 Eros (largest NEA)	65×10^6 yr ago	Fires, dust, darkness; atmosphere/ocean chemical changes; large temperature swings	Half of species extinct
Small $R > 1$ km	~ 500 NEAs	$\sim 300,000$ yr ago	Global dusty atmosphere for months	Photosynthesis interrupted; individuals die but few species extinct; civilization threatened
Very small $R > 100$ m	Tonguska event	91 yr ago	Major local effects; minor hemispheric effects; dusty atmosphere	Newspaper headlines; romantic sunsets increase birth rate

The effects of an impact on life depend in a qualitative way on the impact energy. The smallest space debris to hit Earth's atmosphere is slowed to benign speeds by gas drag or vaporized before it hits the ground. The largest impactors can melt a planet's crust and eliminate life entirely. Strong iron impactors ranging in size from that of a car to that of a house may hit the ground at high velocity, killing living beings in their path. The rocky bolide that exploded over Tonguska, Siberia, in 1908 was about the size of a football field; it produced a blast wave that knocked over trees tens of kilometres away. An impactor a few kilometres in size would throw enough dust into the upper atmosphere to substantially darken the sky for much of a growing season¹⁶; the threat of such an impactor to human food supplies has led NASA to initiate a programme to detect all near-Earth asteroids (NEAs) larger than about 1 km. Mass extinctions (such as that at the K/T boundary; see text) result from even larger impacts, which load the atmosphere with dust and chemicals (from vapour and pulverized matter originating in both the impactor and the crater ejecta); radiation from high-velocity ejecta re-entering the atmosphere may cause global fires. Even larger impacts fill the atmosphere with enough hot ejecta and greenhouse gases to vaporize part or all of the planet's oceans²⁰. Indeed, phylogenetic evidence implies that the last common ancestor of all life on Earth was a thermophilic prokaryote, which would have been most capable of surviving such a scalding impact. Even larger impacts would destroy all life on the planet, although it is possible that some organisms could survive in meteoroids ejected by the impact, and subsequently re-establish themselves on the planet (or another planet orbiting the same star) by a fortuitously gentle return after conditions on the planet had improved.



NATURAL HISTORY MUSEUM, LONDON

photochemical reactions create organic molecules. Analogous processes may have occurred within Earth's early atmosphere.

Interstellar probes for *in situ* exploration of extrasolar planets require substantial technological advances. Even with gravity assists by the planets and the Sun, payloads sent with current rocket propulsion systems would require thousands of years to reach the nearest stars. Thus at present, interstellar travel remains in the realm of science fiction, but considering the advances of the past millennium, voyages over such vast distances may become practical in the coming centuries.

Another approach to the detection of habitable planets is to search for signals that have been sent, intentionally or otherwise, by inhabitants of those worlds. This is being done by the Search for ExtraTerrestrial Intelligence (SETI) (see Box 1).

Although we are not yet able to reach the stars, we are nonetheless entering a golden age of extrasolar planetary study by means of telescopic observations. All of the known extrasolar planets have been identified indirectly, through the gravitational force that they exert on their star, and all have been found during the 1990s. The first two extrasolar planets to be discovered orbit a rapidly spinning neutron star, which emits a substantial fraction of its luminosity as X-rays. The extrasolar planets known to orbit main-sequence stars are each more massive than Saturn. They were discovered using the Doppler technique, which measures changes in the radial velocity of the star in response to the planet's gravitational tug.

Various techniques should increase our knowledge of extrasolar planets in the coming decades. Doppler studies should continue to find giant extrasolar planets⁶: the current precision of this technique is sufficient to detect Jupiter-mass planets in Jupiter-like orbits, and Uranus-mass planets orbiting very close to their stars. Better precision may eventually lead to identification of smaller and more distant planets, but turbulence and other variability of stellar atmospheres will make detection of Earth analogues using the Doppler technique impractical if not impossible.

Planets may also be detected through the wobble that they induce in the motion of their stars projected into the plane of the sky. This astrometric technique is most sensitive to massive planets orbiting stars that are relatively close to the Earth. Because the star's motion is detectable in two dimensions, a better estimate of the planet's mass can be obtained than by using radial velocities. No astrometric claim of detecting an extrasolar planet has yet been confirmed, but technological advances (especially the development of optical and infrared interferometry) suggest that Jupiter-mass

(and possibly smaller) planets will be detected from the ground using this technique within the next decade. Even higher precision is likely to be achievable from spacecraft observations. The ultimate limit to astrometric precision is likely to be differences between the positions of a star's centre of mass and its centre of light, which result from 'starspots' and other variations in brightness.

Earth-sized extrasolar planets?

If the Earth lies in or near the orbital plane of an extrasolar planet, that planet passes in front of the disk of its star once each orbit as viewed from Earth. Precise photometry can reveal such transits, which can be distinguished from rotationally modulated 'starspots' and intrinsic stellar variability by their periodicity, and would provide the size and orbital period of the detected planet. Scintillation in, and variability of, the Earth's atmosphere limit photometric precision to roughly one-thousandth of a magnitude, allowing detection from the ground of transits by Jupiter-sized planets but not by Earth-sized planets¹⁵. Far greater precision is achievable above the atmosphere, with planets as small as Earth likely to be detectable¹⁶. This technique has the greatest potential for detecting Earth analogues within the next ten years.

Distant planets are very faint objects which are located near much brighter objects (the star or stars that they orbit), and thus they are extremely difficult to image. Efforts are currently underway to image giant planets directly from the ground using adaptive optics (which adjusts telescope mirrors to compensate for the variability of Earth's atmosphere) and coronagraphs (which block the light from the star that the planet orbits). NASA and the European Space Agency, ESA, are both currently designing space missions for the second decade of the twenty-first century that will use interferometry and nulling to search for Earth-like planets around nearby stars¹⁷. Assuming such planets are detected, spectroscopic investigations of their atmospheres, to search for gases such as oxygen, ozone and methane, will follow. Obtaining resolved images of the disks of extrasolar Earth-like planets requires substantially better optics and much greater light-gathering capability than available at present, but unlike interstellar travel, such distant resolved images should be achievable within the twenty-first century.

Observations of planetary formation provide information on properties of planets as a class, and typical fluxes of material (impactors) to which Earth-like planets are exposed. The infrared region of the spectrum is likely to provide the greatest information about star and planet formation in the coming decades, because planetary

regions radiate most of their energy in the mid-infrared and dusty interstellar clouds transmit far more scattered starlight in the near-infrared than in the visible range. Advances in infrared detectors, interferometry and new large telescopes (on the ground with adaptive optics, and in space¹⁸) will provide data that are vastly superior to those available today.

Prospects

Predictions have a miserable success rate, and forecasts centuries into the future tend to be particularly conservative. In part, this reflects deficiencies in the human imagination. However, the average rate of change greatly exceeds the median rate. Consider the implication of joining the Galactic club of civilizations a billion years more advanced than our own (assuming such a community exists!). Such a revelation would lead to changes far more fundamental than the invention of movable type, the industrial revolution and the information age have brought to us within the past millennium.

We still do not know whether Earth-like planets on which liquid water flows are rare, are usual for solar-type stars or have intermediate abundances. Nonetheless, I personally believe that, as there are billions of Sun-like stars in our Galaxy, planets with liquid water oceans lasting for long periods of time are common enough to ensure that if we are the only advanced life form in our part of the Galaxy, biological factors are much more likely to be the principal limiting factor than are astronomical causes.

Jack J. Lissauer is at the Space Science Division, MS 245-3, NASA Ames Research Center, Moffett Field, California 94035, USA.

e-mail: lissauer@ringside.arc.nasa.gov

- Lewis, J. S. *Worlds Without End: The Exploration of Planets Known and Unknown* (Helix, Reading, Massachusetts, 1998).
- Joshi, M. M., Haberle, R. M. & Reynolds, R. T. *Icarus* **129**, 450–465 (1997).
- Lissauer, J. J. *Rev. Mod. Phys.* **71**, 835–845 (1999).
- Weidenschilling, S. J. & Cuzzi, J. N. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1031–1060 (Univ. Arizona Press, Tucson, 1993).
- Safronov, V. S. *Evolution of the Protoplanetary Cloud and Formation of the Earth and Planets* (Nauka Press, Moscow, 1969) (in Russian); Publication TTF-677, NASA, 1972 (English transl.).
- <http://www.physics.sfsu.edu/~gmarcy/planetsearch/planetsearch.html>
- Wetherill, G. W. *Astrophys. Space Sci.* **212**, 23–32 (1994).
- Lissauer, J. J. *Icarus* **114**, 217–236 (1995).
- Pollack, J. B. *et al.* *Icarus* **124**, 62–85 (1996).
- Chapman, C. R. & Morrison, D. *Nature* **367**, 33–40 (1994).
- McKay, C. P. *Orig. Life Evol. Biosphere* **27**, 263–289 (1997).
- <http://mpfwww.jpl.nasa.gov/>
- http://www.jpl.nasa.gov/ice_fire/europao.htm
- <http://www.jpl.nasa.gov/cassini/>
- <http://web99.arc.nasa.gov/~mars/vulcan/>
- <http://www.kepler.arc.nasa.gov/>
- <http://astro1.bnscl.rla.ac.uk:80/darwin/>
- <http://ngst.gsfc.nasa.gov/>
- Toon, O. B., Zahnle, K., Morrison, D., Turco, R. P. & Covey, C. *Rev. Geophys.* **35**, 41–78 (1997).
- Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. *Nature* **342**, 139–142 (1989).
- <http://www.seti.org/>

Acknowledgements. I thank J. Cuzzi, F. Drake and G. Marcy for insightful discussions.

Genetics and general cognitive ability

Robert Plomin

General cognitive ability (g), often referred to as 'general intelligence', predicts social outcomes such as educational and occupational levels far better than any other behavioural trait. g is one of the most heritable behavioural traits, and genes that contribute to the heritability of g will certainly be identified. What are the scientific and social implications of finding genes associated with g ?

During the past three decades, the behavioural sciences have emerged from an era of strict environmental explanations for differences in behaviour to a more balanced view that recognizes the importance of nature (genetics) as well as nurture (environment). This shift occurred first for behavioural disorders, including rare disorders such as autism (which has an incidence of 1 per 1,000 population), more common disorders such as schizophrenia (1 in 100), and very common disorders such as reading disability (1 in 50). More recently, it has become increasingly accepted that genetic variation makes an important contribution to differences among individuals in the normal range of behaviour as well as for abnormal behaviour. Moreover, many behavioural disorders, especially common ones, may represent variation at the extremes of the same genetic and environmental factors that are responsible for variation within the normal range. For example, disorders such as reading disability may not be due to genetic variants that specifically influence the disorder. Rather, the same genes that contribute to the normal range of individual differences in reading ability may be responsible for reading disability (Fig. 1). This view, known as the quantitative trait locus (QTL) perspective (see Box 1), has important implications for the search for the genes responsible for behaviour because such genes will individually have small effects; this will make them more difficult to find than genes that have major effects¹.

General cognitive ability

For historical and political reasons, one quantitative trait in particular is highly controversial. This is general cognitive ability, which has a normal distribution in the population from a low end of mild mental handicap to a high end of gifted individuals. Diverse measures of cognitive abilities — such as spatial ability, verbal ability, information processing speed and

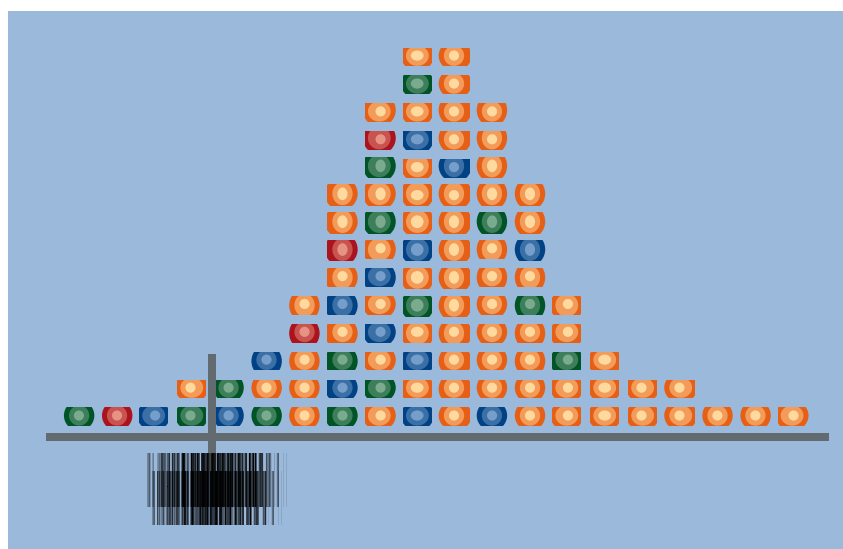


Figure 1 Quantitative trait locus (QTL) perspective on complex traits. Differences among individuals for most quantitative or complex traits such as reading ability are distributed as a normal bell-shaped curve. Multiple genes influence complex traits as probabilistic propensities rather than predetermined programmes. Here the different genetic make-up of individuals with respect to two hypothetical genes involved in reading ability is shown for 100 individuals (each person is represented by an oval), with five of these individuals (those on the extreme left) receiving a diagnosis of reading disability. The green ovals indicate that the individual has the disabling variant of one gene and blue ovals denote the disabling variant of the other gene. Neither gene is necessary or sufficient for low scores, even for individuals who have disabling variants of both genes (red ovals). This QTL perspective suggests that genes associated with common disorders such as reading disability may represent the quantitative extreme of the same genes that are responsible for variation throughout the population.

memory — correlate substantially with each other, and general cognitive ability (g) is what these diverse measures have in common (see Box 2). Clearly there is more to cognition than g — although g explains about 40 per cent of the variance among such tests, most of the variance of a particular test is independent of g .

There is a wide gap between what lay people (including scientists in other fields) believe about intelligence and intelligence testing, and what the professional behavioural scientist believes. Most notably,

lay people often read in the popular press that the assessment of intelligence is circular — intelligence is what intelligence tests assess. On the contrary, g is one of the most reliable and valid measures in the behavioural domain; its long-term stability after childhood is greater than for any other behavioural trait, and it predicts important social outcomes such as educational and occupational levels far better than any other trait². Although a few critics remain³, g is widely accepted by experts⁴. But it is less clear what g is: is it due to a single general process

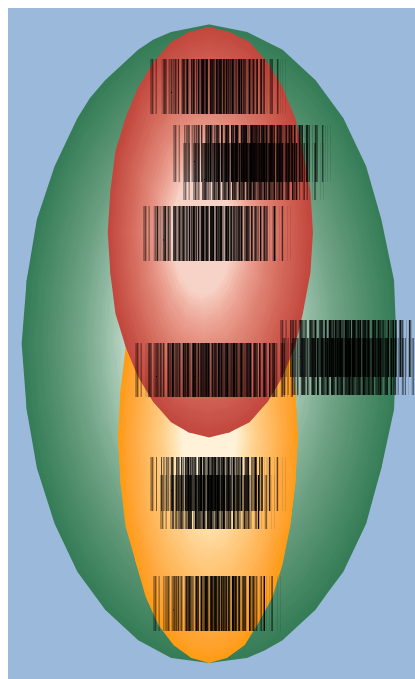


Figure 2 Functional genomics includes all levels of analysis from molecular biology to psychology. The higher levels of analysis can be referred to as behavioural genomics in order to emphasize the importance of top-down analyses of pathways between genes and behaviour.

such as high-level strategies called executive function or speed of information processing, or does it represent a concatenation of more specific cognitive processes?^{5,6}

The concept of a genetic contribution to *g* has provoked much controversy, especially following the publication in 1994 of *The Bell Curve* by Herrnstein and Murray⁷. (This book in fact scarcely touches on genetics and does not view genetic evidence as crucial to its arguments.) The first half of the book shows, like many other studies, that *g* is related to educational and social outcomes, but the second half attempts to argue that certain right-wing policies follow from these findings. However, as discussed later in this article, public policy does not necessarily follow from scientific findings, and it would be possible to argue in just the opposite direction from Herrnstein and Murray. Nonetheless, there is considerable consensus among scientists — even those who are not geneticists — that *g* is substantially heritable^{6,8,9}. Indeed, there are more studies addressing the genetics of *g* than of any other human characteristic, including studies of more than 8,000 parent–offspring pairs, 25,000 pairs of siblings, 10,000 twin pairs and hundreds of adoptive families, all of which indicate that genetic factors contribute significantly to *g*^{10,11}.

Estimates of the size of the genetic effect, which population geneticists call ‘heritability’, vary from 40 to 80 per cent, but estimates based on the entire body of data make it about 50 per cent, indicating that genes account for about half of the variance in *g*. When the data are sorted by age, heritability is found to increase from about 20 per cent in infancy, to about 40 per cent in childhood, to 60 per cent or greater later in life¹², even for individuals over 80 years old¹³. This increase in heritability throughout lifespan is interesting, because it is counterintuitive to the effects of Shakespeare’s ‘slings and arrows of outrageous fortune’ accumulating over time. It may be that heritability increases because individuals seek out and create environments correlated with their genetic propensities.

Most of the genetic variance for *g* is additive, that is, the effects of the individual genes seem simply to add up rather than there being interactions between the genes. The additivity of most genetic effects on *g* may be because there is greater assortative mating (non-random mating) for *g* than for any other behavioural trait. In other words, bright women are likely to mate with bright men and the outcome of this dual effect is that their offspring are likely to be brighter on average than would be expected if mating were at random, thus spreading out the distribution of *g* in the population.

The data that provide evidence for a genetic effect on *g* also provide the best available evidence for the importance of environmental factors that are independent of genetics. Environment clearly is important, as indicated by the steady rise in IQ scores during the past several generations, which would seem too short a time to be explained by genetics¹⁴, and by studies in which children from abusive families show gains in IQ when adopted¹⁵.

Recent findings and new directions

Genetic research has moved beyond the rudimentary questions of whether, and to what extent, genetic differences are important in the origins of individual differences in *g*. These new findings inform the scientific and social implications discussed later.

It is often not appreciated that genetically sensitive designs, such as twin and adoption studies, that recognize the importance of both nature and nurture are uniquely well suited to the investigation of environmental influences. Indeed, one of the most important discoveries about environmental influences on *g* has come from such genetic research. The ‘nurture assumption’¹⁶ — that the home is the most important part of the child’s environment — implies that children growing up in the same home should be similar to one another because they share these environmental influences. When genetic resemblance is taken into account, such shared environmental influences that contribute to the resemblance of

family members for *g* are important in childhood, accounting for about a quarter of the variance, but they are not important after adolescence. In other words, non-shared environmental factors that make children in the same family different (such as differences in parental treatment, differences in school experiences and different experiences with peers) provide the long-term consequences of environmental influence for *g*. This finding coincides with similar findings for other quantitative traits and indicates the need to re-examine the nurture assumption¹⁶.

Two other examples of recent genetic findings about the environment are that environmental influences may override genetic effects in families of low socioeconomic status¹⁷, and that genetic factors contribute to individuals’ interactions with their environment¹⁸. The former finding highlights that heritability estimates are not absolute but depend on the environment in which they are measured. The latter observation, called ‘genotype–environment correlation’, indicates that genetic influences on abilities can best be thought of as

Box 1 The source of genetic variation

Quantitative traits are those characteristics, such as height or ‘intelligence’, that are found as a continuum of values within a population rather than as the discrete alternative inherited character states familiar to most people from a schoolroom acquaintance with Mendel’s peas. They are due to the combined effects of a number of different genes (each of which will, of course, be inherited according to the rules of mendelian genetics), and usually also have a considerable environmental input to the final outcome. This final outcome is known as the phenotype. A quantitative trait locus (QTL) refers to a gene that contributes to a quantitative trait. A locus is the technical name in genetics for the position on the chromosome at which a gene for a particular characteristic is located. In any outbreeding population, such as humans, many genes are present in the population in a number of variant forms (technically known as alleles). By inheritance from their mother and father each person carries two copies (that is, two alleles, which may be different or identical) of the genes corresponding to most loci (those on the sex chromosomes excepted). The genetic variation within the human population is due to the immense number of different combinations of alleles possible, given the tens of thousands of different loci in the human genome. These different combinations give each individual human being (except identical twins) a unique genetic make-up or genotype, even though all humans share the same set of loci.

appetites rather than aptitudes in the sense that genetic propensities propel individuals towards evoking, selecting and constructing experiences that are correlated with their genetic propensities. Bright children select (and are selected by) peers and educational programmes that foster their abilities. They read and think more.

Work on genetic influences on intelligence has up to now focused on *g*. Much less is known about the genetic and environmental origins of individual differences in specific cognitive abilities such as spatial ability, verbal ability, memory and processing speed. Specific cognitive abilities show substantial genetic influence, although less than for *g*¹⁹. To what extent do different sets of genes affect these different abilities? A technique called multivariate genetic analysis examines covariance among specific cognitive abilities and yields a statistic called the 'genetic correlation', which is the extent to which genetic effects on one trait correlate with genetic effects on another trait independently of the heritability of the two traits. That is, although individual cognitive abilities are moderately heritable, the genetic correlations between them could be anywhere from 0, indicating that a completely different set of genes influences each ability, to 1, indicating that the same genes influence a variety of cognitive abilities. Multivariate genetic analyses have consistently found that genetic correlations among specific cognitive abilities are very high — close to 1 (ref. 20).

Implications for cognition theory

These results have major implications for current theories of how the brain works. According to one theory, the brain works in a modular fashion — the various cognitive processes are specific and independent of each other. Implicit in this perspective is a bottom-up reductionistic view of genetics in which individual modules such as spatial visualization are the targets of gene action. However, the findings from multivariate genetic analyses suggest a top-down view, in which genetic effects operate primarily on *g*. Given that the brain has evolved to learn from a variety of experiences and to solve a variety of problems, perhaps it makes sense that it would function holistically.

Nevertheless, finding genetic correlations near 1.0 does not prove that genetic effects are limited to a single general cognitive process. Another alternative is that specific cognitive abilities might tap many of the same modular processes that are each affected by different sets of genes. This hypothesis could be tested by multivariate genetic research on measures of modular processes, for example determining which areas of the brain are active in response to a particular task. Such procedures can now be done using neuroimaging techniques, and

Box 2 General cognitive ability (*g*)

General cognitive ability (*g*) is best captured by a technique called factor analysis in which a composite score is created that represents what diverse measures of cognitive abilities have in common (the operational definition of *g*). That is, tests of spatial ability correlate moderately with tests of verbal ability, and memory tests correlate with spatial and verbal tests, although more modestly. Factor analysis (an unrotated first principal component) creates a composite that weights each test by its overall correlation with all other tests. This vector of weights, called factor loadings, correlates with the heritabilities of the tests, that is, high *g*-loaded tests are the most heritable. In addition, *g* loadings correlate with other biological and psychological processes, using a technique called the 'method of correlated vectors', which supports the validity of *g*².

Although *g* is one definition of intelligence, intelligence has so many other connotations that it has been suggested that the word be avoided in scientific discussion⁵. Intelligence tests such as the individually administered Wechsler tests, which are widely used for clinical purposes, assess diverse abilities such as spatial ability (for example, making multicoloured blocks match a two-dimensional design), verbal ability (for example, vocabulary), speed of processing (for example, matching digits to symbols), memory (for example, memorizing a sequence of digits) and reasoning (for example, identifying what is missing in a picture). Rather than weighting these tests by their contribution to *g*, intelligence tests weight each test equally by summing them to create a composite score known as IQ, which is standardized to have a mean of 100 and a standard deviation of 15. Nonetheless, IQ scores correlate highly ($r \approx 0.80$) with *g* scores derived from factor analysis.

these tests could investigate whether genetic correlations are also close to 1 between these modular measures and general processes²¹. As with many unanswered questions about genetics and cognitive abilities, clearer answers will emerge when specific genes are identified. In this case, the question is whether genes associated with *g* are associated with a single general cognitive process, or with most modular processes, or with specific subsets of modular processes.

Finding the genes for *g*

The most far-reaching implications for science, and perhaps for society, will come from identifying genes responsible for the heritability of *g* — not rare single-gene mutations that cause mental retardation, but QTLs that contribute probabilistically to individual differences in the normal variation in *g*.

Such loci are 'polymorphic'; that is, there are at least two, and often many more, variant forms of the gene in the population. These variants originally arise by mutations that change the actual coding sequence of a gene, thus producing a slightly different form of the encoded protein, or that affect the regulatory parts of a gene, thus affecting when and where the gene is switched on and protein produced. Both types of variation could contribute to the heritability of *g*.

Presumably, genes that are active in the brain ('expressed' in the brain in genetic terminology) are involved in specifying *g*, but with 30,000 or so genes known to be expressed in the brain this hardly helps to narrow the field. A small handful of genetic associations with behaviour have been found so far. The first definite associations of QTLs with behaviour have emerged in the area of cognitive disabilities, namely, dementia (an association with the

apolipoprotein E gene) and reading disability (which has been linked to a region on the short arm of chromosome 6). Associations of *g* with identified polymorphic segments of DNA on the genetic map (DNA markers) have begun to be reported²². As a result of the progress made in mapping the human genome, it is now becoming feasible to carry out genome scans using association approaches involving several thousand closely spaced DNA markers. These have the power to detect and locate the kind of genes that are likely to contribute to *g*, that is, multiple genes of small effect. The initial results of a systematic genome scan of thousands of DNA markers reported several replicated QTL associations²³. The massive effort needed to genotype thousands of DNA markers for large numbers of subjects is daunting and replication is needed to eliminate false positive results. However, optimism about this approach has been fuelled by the promise of 'SNPs on chips' — single nucleotide polymorphisms (SNPs) formatted as microarrays of oligonucleotide primers on solid substrates that can quickly genotype thousands of DNA markers for an individual. So far, such microarrays have been most useful in studies of gene expression and there remain technical difficulties to be overcome before they can be routinely used to genotype SNPs for large samples.

I have no doubt that genes associated with *g* will be identified, although how much of the genetic variance will be accounted for by individual genes is uncertain. This is because the magnitude of the effects of genes in multiple gene systems is not yet known for *g* or for any other trait or disorder controlled by a number of different genes (also known as complex traits and polygenic disorders). It is likely that the average size of effect of each individual gene is small for complex traits —

perhaps individual genes on average will account for less than 1 per cent of the variance, with few effect sizes greater than 5 per cent, and a long tail of small effects extending to genes of such small effect that they may never be detected. If genes interact with each other they will be more difficult to identify because interactive combinations of genes would need to be found rather than individual genes. Fortunately, genetic effects on *g* seem to be largely additive. Despite the formidable challenges of trying to find genes of small effect, I predict that most of the heritability of *g* will be accounted for eventually by specific genes, even if hundreds of genes are needed to do it.

So what are the scientific and social implications of finding genes that influence *g*?

Implications for science

One of the first tasks is to localize the specific DNA differences responsible for associations between DNA markers and *g*. Finding the locations of the genes involved will be greatly aided by several developments. The first will be the completion of the entire DNA sequence of the human genome, which is expected from the Human Genome Project during the next two years; the second is the recent intensive effort to identify hundreds of thousands of DNA variations among individuals, which will make it possible to pinpoint functional variants; and the third is the effort underway to map patterns of gene expression, which will indicate which genes are expressed in any given brain region.

The ultimate scientific goal is not just finding the genes but understanding how they function, and this is an even more challenging task. Functional genomics, as this aspect of genetics is called, is usually discussed in terms of molecular biological analyses of cell function. For example, studies of coordinated spatial and temporal patterns of gene expression using DNA chips containing detectors for thousands of genes — like functional imaging at a cellular level — will make major contributions to the study of gene function. However, other levels of analysis are also important in understanding how genes work. A behavioural level of analysis will also contribute to functional genomics, for example, by means of psychological theories of cognitive processing⁶ and by investigating interactions and correlations between individuals and their environment²⁴. For instance, psychological theories suggest how different components of information processing are related and the role of genes in these cognitive systems can be examined. Such top-down strategies can yield just as important information as a bottom-up molecular approach in which the products of these genes are studied at a cellular level of analysis. As an antidote to the

tendency to define functional genomics at the cellular level of analysis, the phrase ‘behavioural genomics’ has been proposed²⁵ (see Fig. 2). Indeed, behavioural genomics may pay off more quickly than other levels of analysis in terms of prediction, diagnosis, therapy and intervention in relation to behavioural disorders and normal behavioural variation.

The brain is clearly where bottom-up molecular levels of analysis will eventually meet top-down behavioural analysis. Studies of brain functioning, as assessed by neuroimaging for example, will foster this integration. For humans, the expense of neuroimaging restricts sample sizes and, for this reason, differences between individuals are rarely considered. Mouse models will be valuable, especially given current large-scale behavioural screens of mice treated with mutagens. Finding genes associated with *g* and other cognitive abilities and disabilities in humans will provide discrete windows through which brain pathways leading from genes to complex cognitive processes of learning and memory can be observed using animal models. Although learning and memory are the focus of much animal research in cognitive neuroscience, commonalities across cognitive processes that are indicative of *g* have not yet been explored using animal models. The main focus at present is the study of synaptic connections between brain neurons and how they may be altered as an animal (or a person) learns or lays down a memory. The multivariate genetic results mentioned earlier lead to the hypothesis that, even at the cellular level, for example in connection with synaptic alterability, most genes will have broad effects on cognitive functioning rather than isolated effects on individual modules. That is, the same genes will affect many different brain structures and cellular processes. Although research on synaptic mechanisms of learning and memory is leading the way in genetic research in cognitive neuroscience²⁶, investigators have used gene knock-outs, in which a gene is altered so that it is no longer expressed, rather than studying naturally occurring genetic variation that might underlie individual differences in learning, memory and cognitive abilities. It is possible, but not inevitable, that a gene identified as being involved in learning and memory in gene knock-out experiments may also turn out to contribute to individual differences in *g* when naturally occurring differences in the gene are identified.

The scientific impact of finding genes associated with *g* will not be limited to cognitive neuroscience — it will affect all aspects of behavioural research. Perhaps some day behavioural and social scientists will routinely collect DNA using cheek swabs (where no blood is needed) in order to investigate, or at least control for, genes associated with *g*, as

is happening now in research on dementia and its only known risk factor, the gene for apolipoprotein E. Even if hundreds of genes contribute to the heritability of *g*, finding genes associated with *g* will make it possible to investigate long-standing scientific issues with much greater precision. For example, in relation to the finding that heritability of *g* increases during development, are there additional genes associated with *g* later in life, or do the same genes have greater effects as time progresses? What are the mechanisms by which gene–environment interactions and correlations emerge? Will the same genes affect different cognitive abilities and modular measures of brain function?

In terms of treatment-related research, finding genes associated with *g* is likely to lead to gene-based diagnoses and treatment programmes for mild mental retardation, and clarification of its overlap with learning disabilities. Gene-based classification of disorders may bear little resemblance to our current symptom-based diagnostic systems. Indeed, from a QTL perspective, common disorders may be the quantitative extreme of normal genetic variation rather than qualitatively different. The most exciting prospect is for secondary prevention — if DNA analysis can be used to predict genetic risk for an individual, this might offer the hope of intervention before disorders create cascades of complications. The exemplar is phenylketonuria (PKU), a disorder that is due to a defect in a single gene. This leads to mental retardation unless it is detected early in life and dietary intervention is used to ameliorate its effects on the developing brain.

Perhaps the greatest implication for science is that the functional genomics of *g* and other complex traits will serve as an integrating force across diverse disciplines, with DNA as the common denominator, opening up new scientific horizons for understanding behaviour.

Implications for society

At the outset, it should be emphasized that no policies necessarily follow from finding genes associated with *g*, because policy involves values. For example, finding genes associated with *g* does not mean that we ought to put all of our resources into educating the brightest children. Depending on our values, we might worry more about children falling off the low end of the normal distribution in an increasingly technological society, and decide to devote more public resources to those in danger of being left behind. For example, all citizens need to be computer literate so that they will not be left on the shore while everyone else is surfing the World-Wide Web. There is much room for values here because these issues involve a complex balancing act among the rights and responsibilities of society, parents and

children. The only thing that seems completely clear is that nothing will be gained by ignoring the issue and pretending that *g* does not have a significant genetic component. As is always the case, advances in science create new challenges; we should be alert to such possibilities so that we can position ourselves to maximize the gains and minimize the pains of new discoveries. Recommended reading is an analysis of the ethics of the genetics of *g* written by an ethicist and a molecular geneticist²⁷.

Will a DNA chip for *g* make the 1997 science fiction film *GATTACA*, in which individuals are selected for education and employment on the basis of their DNA, come true? After childhood, no DNA chip can predict *g* as well as an IQ test given to that individual, because IQ tests measure the consequences of environmental influences as well as genetic ones. I think it is more likely that educators and employers interested in *g* will continue to use IQ tests and achievement tests to select individuals on the basis of what they can do, rather than using a DNA chip to estimate what they could have done. However, such a DNA chip might be used in education to consider how far children are fulfilling their genetic potential or to prescribe different training programmes.

DNA chips for *g* might be used for prenatal testing, for example, in the selection of embryos for *in vitro* fertilization. But this seems unlikely, because few embryos are available to choose from and there are many important genetic diseases to screen out. What about parents who want to use DNA chips for *g* in order to select egg or sperm donors, because such a chip might provide better estimates of genetic potential than phenotypic tests of the donors? Is it possible that there are parents who would use DNA chips for *g* prenatally for eugenic purposes?

Will DNA chips for *g* be used for postnatal screening to enable interventions that avoid risks or enhance strengths? For decades we have screened newborns for PKU because a relatively simple dietary intervention exists that prevents its damage to the developing brain. If similarly low-tech and inexpensive interventions such as dietary changes could make a difference for some *g* genotypes, parents might want to take advantage of them even if the QTL accounts for only a small amount of variance. Expensive high-tech genetic engineering in regard to behavioural traits is unlikely to happen for a long time — it is proving very difficult even for the single-gene disorders, and it will be many orders of magnitude more difficult and less effective for complex traits influenced by many genes.

A more general concern involves group differences, such as average differences between classes and ethnic groups. As genes are found that are associated with differences among individuals within groups, the genes

will inevitably be used to make comparisons between groups. Although such comparisons will not be straightforward because the causes of individual differences within groups are not necessarily related to the causes of average differences between groups, the societal implications of such research need to be anticipated. To keep this issue in perspective, it should be emphasized that average differences in *g* between groups are small compared with the range of individual differences within groups. Perhaps people will become less preoccupied with average differences between groups when DNA chips make it possible to focus on individuals.

Another general concern is that knowledge about the importance of genetics might change attitudes — for example, attitudes of parents about the malleability of their children's cognitive ability. If there are parents who do not recognize genetic limits to their children's ability, it might actually be useful for them to have a more realistic view, so that their children's failures are not interpreted as simple motivational failures. Do parents matter? They do indeed. And not just because of their genes. Although the genetic research discussed earlier indicates that parents do not mould their children environmentally to be similar to them in terms of *g*, genetic research on *g* is mute about much that parents offer their children as teachers and models independent of the parents' *g*. Moreover, as Judith Harris concludes in her book, *The Nurture Assumption*: "We may not hold their tomorrows in our hands but we surely hold their todays, and we have the power to make their todays very miserable."¹⁶

The most general fear is that finding genes associated with *g* will undermine support for social programmes because it will legitimate social inequality as 'natural'. The unwelcome truth is that equal opportunity will not produce equality of outcome because people differ in *g* in part for genetic reasons. When the US founding fathers declared that all men are created equal they did not mean that all people are identical, but rather that they should be equal before the law. Democracy is needed to ensure that all people are treated equally despite their differences. On the other hand, finding heritability or even specific genes associated with *g* does not imply that *g* is immutable. Indeed, genetic research provides the best available evidence that non-genetic factors are important in the development of individual differences in *g*. PKU provides an example that even a single gene that causes mental retardation can be ameliorated environmentally.

'There is no gene for the human spirit' is the subtitle of the film *GATTACA*. It embodies the fear lurking in the shadows that finding genes associated with *g* will limit our freedom and our free will. In large part such

fears involve misunderstandings about how genes affect complex traits like *g*²⁸. Finding genes associated with *g* will not open a door to Huxley's brave new world where babies are engineered to be alphas, betas and gammas. The balance of risks and benefits to society of DNA chips for *g* is not clear — each of the problems identified in this section could also be viewed as a potential benefit, depending on one's values. What is clear is that basic science has much to gain from functional genomic studies of brain functions related to learning and memory. We need to be cautious and to think about societal implications and ethical issues. But there is also much to celebrate here in terms of the increased potential for understanding our species' unparalleled ability to think and learn.

Robert Plomin is at the Institute of Psychiatry, King's College London, DeCrespigny Park, Denmark Hill, London SE5 8AF, UK.

e-mail: r.plomin@iop.kcl.ac.uk

1. Plomin, R., Owen, M. J. & McGuffin, P. *Science* **264**, 1733–1739 (1994).
2. Gottfredson, L. S. *Intelligence* **24**, 13–23 (1997).
3. Gould, S. J. *The Mismeasure of Man* (W. W. Norton, New York, 1996).
4. Carroll, J. B. *Intelligence* **21**, 121–134 (1995).
5. Jensen, A. R. *The g factor: The Science of Mental Ability* (Praeger, Westport, 1998).
6. Mackintosh, N. J. *IQ and Human Intelligence* (Oxford University Press, Oxford, 1998).
7. Herrnstein, R. J. & Murray, C. *The Bell Curve: Intelligence and Class Structure in American Life* (The Free Press, New York, 1994).
8. Brody, N. *Intelligence* (Academic Press, New York, 1992).
9. Snyderman, M. & Rothman, S. *The IQ Controversy, the Media and Publication* (Transaction, New Brunswick, NJ, 1988).
10. Bouchard, T. J. Jr & McGue, M. *Science* **212**, 1055–1059 (1981).
11. Plomin, R., DeFries, J. C., McClearn, G. E. & Rutter, M. *Behavioral Genetics* (W. H. Freeman, New York, 1997).
12. McGue, M., Bouchard, T. J., Iacono, W. G. & Lykken, D. T. in *Nature, Nurture, and Psychology* (eds Plomin, R. & McClearn, G. E.) 59–76 (American Psychological Association, Washington, DC, 1993).
13. McClearn, G. E. *et al. Science* **276**, 1560–1563 (1997).
14. Flynn, J. *Am. Psychol.* **54**, 5–20 (1999).
15. Duyme, M., Dumaret, A.-C. & Tomkiewicz, S. *Proc. Natl Acad. Sci. USA* **96**, 8790–8794 (1999).
16. Harris, J. R. *The Nurture Assumption: Why Children Turn Out the Way They Do* (The Free Press, New York, 1998).
17. Rowe, D. C. & van den Oord, E. J. C. G. *Child Dev.* (in the press).
18. Plomin, R. *Genetics and Experience: The Interplay Between Nature and Nurture* (Sage Publications, Thousand Oaks, CA, 1994).
19. Plomin, R. & DeFries, J. C. *Sci. Am.* **May**, 62–69 (1998).
20. Petrill, S. A. *Curr. Directions Psychol. Sci.* **6**, 96–99 (1997).
21. Kosslyn, S. & Plomin, R. in *Psychiatric Neuroimaging Strategies: Research and Clinical Applications* (eds Dougherty, D., Rauch, S. L. & Rosenbaum, J. F.) (American Psychiatric Press, Washington, DC, in the press).
22. Chorney, M. J. *et al. Psych. Sci.* **9**, 1–8 (1998).
23. Fisher, P. I. *et al. Hum. Mol. Genet.* **8**, 915–922 (1999).
24. Plomin, R. & Rutter, M. *Child Dev.* **69**, 1221–1240 (1998).
25. Plomin, R. & Crabbe, J. C. *Psychol. Bull.* (in the press).
26. Migaud, M. *et al. Nature* **396**, 433–439 (1998).
27. Newson, A. & Williamson, R. *Bioethics* **13**, 327–342 (1999).
28. Rutter, M. & Plomin, R. *Br. J. Psychiat.* **171**, 209–219 (1997).

Acknowledgements. Preparation of this paper and the QTL research on *g* is supported by a grant from the US National Institute of Child Health and Human Development. The paper profited from review of an earlier version by R. Arden, S. Baron-Cohen, I. Craig, P. Dale, I. J. Deary, J. C. DeFries, L. S. Gottfredson, F. Happé, J. Rich Harris, C. Hughes, D. Lubinski, P. McGuffin, A. Newson, T. G. O'Connor, M. J. Owen, S. A. Petrill, K. J. Saudino, L. A. Thompson and I. D. Waldman.

Tales of the expected

Philip Campbell

Forecasting the future in science is fun but often hopelessly misleading. This publication, commissioned by all the Nature journals, focuses on future developments about which we can be reasonably confident and which will have an impact on the lives of all of us.

As the immunologist Peter Medawar once wrote, wise people may develop expectations about the future, but only the foolish make predictions. X-rays, radioactivity and high-temperature superconductors are just three ways in which physicists have taken themselves and other scientists by delighted surprise. Medawar cited the uncovering of genetic polymorphisms within the human leukocyte antigen system as an unanticipatable simultaneous solution to several biomedical conundrums. The unexpected detection by James Lovelock in the 1960s of chlorofluorocarbons in the marine atmosphere stimulated a transformation in the understanding of the stratospheric ozone layer. And so on.

The unpredictability of experimental progress was neatly summarized in 1926 by J. B. S. Haldane: "In the case of both thunderstorms and fever the clue came from measuring the lengths of mercury columns in glass tubes, but what prophet could have predicted this?" Nevertheless, there has always been the uneasy feeling that governments and legislatures, focused ever more on obtaining benefits from science, expect too much of researchers and of research as a whole in the way of premeditated discovery. That is presumably why learned societies find it necessary to highlight in public what is well known to the cognoscenti: the unanticipatable conjunctions of unrelated developments that have led, time and again, to high peaks of scientific and technological progress. The US National Academy of Sciences has gone so far as to generate a series of case studies in its 'Beyond Discovery' education programme (see www4.nas.edu).

Then again, some of the great achievements of science have flowed from predictions. (Or were they expectations only classed as predictions with hindsight?) When an original way of thinking about a problem leads to a mathematical equation with solutions that represent what have hitherto been mysterious phenomena, that may be a triumph. But excitement runs even higher when the same equation yields solutions that match nothing so far seen. Are they mathematical oddities? The negative-energy solutions in Dirac's equation, which gives a quantum mechanical description of an elementary particle, led him to predict particles of equal mass and opposite charge to the elec-



tron. These were subsequently detected when positrons, the first species of anti-matter, were discovered in cosmic rays. Wolfgang Pauli's physical reasoning led him to conclude with confidence that there must be a very weakly interactive particle and (nearly?) massless particle involved in nuclear decay — subsequently identified in the form of the neutrino. Currently the world awaits the detection of the predicted Higgs boson and — less certainly — 'supersymmetric' partners of all known fundamental particles. (Failure to detect a Higgs particle — a keystone of the currently unshakeable standard model of high-energy physics — at CERN's Large Hadron Collider early next decade is highly improbable but, paradoxically, could be an unpredicted boost for particle physics.)

Scarcity of wise prediction

Such unambiguous predictions, far from foolish, come from within a tightly argued, self-contained theoretical framework based on critical assumptions or hypotheses that may prove false. They stand out among the peaks of science, alongside the unexpected discoveries, and those that were anticipated in principle but still revelatory in actuality, such as the structure of DNA. And not only in physics: witness the analysis that followed Lovelock's work that led atmospheric chemists to predict damage to the ozone layer. It is a shame that, the world being as

messy as it is, such clear-cut predictions — representing scientists at their best, out on an intellectual limb — are so rare.

The more common and foolish predictions about science are usually much more casual and broad ranging, often based on the blinkered preconceptions of even great scientists — see, for example, Rutherford's notorious dismissal of the possibility of nuclear energy. Looking back at predictions 100 years ago, described by John Heilbron and Bill Bynum on page C86, is both entertaining and sobering. *Nature*, hopefully wisely, is straying into predictions only by means of fictions, appearing in our weekly 'Futures' essays. In contrast, this supplement, *Impacts of foreseeable science*, commissioned with the help of the staff of *Nature* and of the six *Nature* monthly journals, is about expectations of what can and might be achieved.

The articles published here are not intended to represent some collective vision, they have not been peer reviewed, nor are they intended to provide balanced surveys. On the contrary: they have been commissioned as representing the scientific hopes and expectations of these particular authors, drawn from among many scientists and professional colleagues who combine a broad overview with an individual outlook. For extra fun, we have interspersed the longer articles with short items about anticipated research relating to a variety of timescales at

OLIVER BURSTON

which natural phenomena are known to occur, from 10^{-43} to (at least) 10^{17} seconds. There is a strong bias in our choice of topics. In the interests of our readership across all disciplines, we have focused on some of the fundamental science that will directly or indirectly touch everybody's lives.

Everyday stuff

In that context, the physics of everyday matter merits a place because, despite such matter's familiarity, fundamental aspects of its behaviour, especially as described by statistical physics, still provide researchers with some of their most difficult intellectual challenges. Condensed matter physics has already led to new electronic technologies that have transformed our lives, not least in developments in information and materials technologies. These are themselves transforming the ways scientists of all disciplines will accomplish computations, as described by Declan Butler on page C67. Who knows what technologies might materialize following a deeper understanding of common and not-so-common transitions between the various phases of matter? Of all the developments discussed in this publication, those are perhaps the most difficult to anticipate in any detail. Philip Ball's article on page C73 thus concentrates on the essential agenda of the subject.

Rather more specific is the expectation that more planets orbiting stars other than the Sun will be discovered. The image of the rising Earth taken by astronauts orbiting the Moon had an impact on many people's views of themselves as members of one population living in perishable but also cherishable conditions. Think, then, of the impact of the image of an Earth-like planet, light-years away, being delivered from a space-based network of interferometers. As Jack Lissauer describes (page C11), such precision is still decades away, but there are likely in the meantime to be other tantalizing planets aplenty.

In highlighting our fragility, the Apollo image was salutary. But we do not understand just how fragile humanity might be, nor what we might do to reduce that fragility. Success at modelling the regional impacts of climate change, for example, still eludes us, let alone societal impacts. We are only now at the stage of developing integrated models incorporating oceans and atmospheres, with necessarily crude representations of uncertain feedbacks involving biota and clouds. Perhaps the most insightful progress will come in a simulatory framework of the sort envisaged by John Schellnhuber on page C19. That will still leave plenty of scope for debate over the application of the precautionary principle as a recipe for societal stagnation or for sensible planetary management.

Biological revolutions

As is now well understood, biologists are at the threshold of an immense collective leap

in knowledge, the impact of which we try to anticipate in some depth. The genes that make up representative genomes across the various kingdoms of life are being identified. The technology is available for mass analyses of gene function and patterns of expression, while the ability to catalogue anybody's genome in a day is now anticipatable, though still lies many years away. Within years we will also have built up libraries of protein structure and function. But the geneticist Eric Lander has characterized molecular and cell biologists as still only approaching a phase of development of their discipline reached by chemists 100 years ago, when the periodic table was first described.

In this supplement we have taken many of those very challenging advances for granted. What will follow? Developmental and evolutionary biologists and palaeontologists will have a wonderful time, as anticipated by Peter Holland on page C41.

But where next for the committed reductionists? Even if one knows all of those elements of an organism, and their individual functions, what could we say about the nature and behaviour of the organism? The answer is: not much, even if the organism consists of a single cell, let alone the 10^{13} cells and 200 or so cell types that make up one of *Homo sapiens sapiens*. This level of analysis takes us to the next step up the ladder whose bottom rung is molecular biology. (The top rung is the science of xenobiology, involving the study of alien life forms. So far we know only one form of life, based on DNA. At this point I will be deliberately foolish and predict the discovery of at least one other form of natural life — not necessarily intelligent — by the end of the next century. But that is a digression.)

Biologists and others have begun to think about systems and networks as such — for example, of neurons and of signalling molecules — but we, or at least Leland Hartwell *et al.*, on page C47, can anticipate such systems analysis of functional, albeit labile, collections of interlinked molecules — or 'modules' — becoming a new norm of biological investigation, crucially aided by the other traditional disciplines and also by computational simulation.

But even ahead of such analysis and understanding, biologists will discover gross characteristics that can be indirectly linked to our molecular constitution. None is likely to be more sensitive, in a societal context, than analysis of cognitive ability and of decision making. Robert Plomin (page C25) and James Nichols and William Newsome (page C35) have their respective views of the power of the sciences to help us understand some of the innate aspects of ability and behaviour. These and other fruits of molecular and organismal biology are readily anticipatable, as are some of the controversies that will ensue. But such applications of fundamental

biology can most favourably be seen as part of the wider picture of future improvement in human health. And, as Barry Bloom describes on page C63, such advances are themselves part of a much bigger picture of development and of debate relating to public health, its priorities and economics.

Public ambivalence towards such advances, and more recently towards undoubtedly positive developments in nutrition (see Gordon Conway and Gary Toenniessen, page C55), are nothing new. As the journalist and sociologist of science Jon Turney has described in his book *Frankenstein's Footsteps: Science, Genetics, and Popular Culture*, the framework for social discussion of contentious science — newspaper coverage, high-profile scientists and hostile representatives of various publics — were all in place at least in the United States and Britain in the 1900s, in time for debates on the chemical basis of life. Aldous Huxley, Julian Huxley and J. B. S. Haldane stimulated such debate in the 1930s as they anticipated much of the developments that have occurred since — and, in *Brave New World*, rather more. The world's first test-tube baby, Louise Brown, stimulated worldwide agitation, and no-one reading this will need to be reminded of Dolly the sheep — a mammal whose unprecedented origin led to new levels of public debate, at all levels of informedness.

Science's contract

In the face of such controversies, there is a lot to be said for an enhanced degree of interaction, genuinely two-way, between science and society. Experience suggests that lack of scientific knowledge by the various publics and stakeholders does not hamper fruitful discussion anything like as much as the reluctance of scientists and technologists to become involved in the sometimes heated, often chaotic and occasionally even physically threatening debates.

Michael Gibbons is someone who has long thought about what is increasingly becoming a commonplace idea: a new contract between science and society (see page C81). His agenda is not idealistic, and is wholly practicable, although his definition of socially robust knowledge may raise some eyebrows. Nevertheless, various types of public participation have been undertaken, for example in the United States, Scandinavia, Britain and France. It is surely time to draw together these experiences and for governments to build on them, if they are to have a hope of bridging the gap between the public's imaginations and the ever swifter and, for some, dismayingly unpredictable development of science.

Philip Campbell is the Editor of Nature and Editor in Chief of Nature publications. For a guide to Nature and the monthly journals, and the relationship between them, see <http://www.nature.com/author/natureguide.html>.

The neurobiology of cognition

M. James Nichols and William T. Newsome

Perhaps the deepest mysteries facing the natural sciences concern the higher functions of the central nervous system. Understanding how the brain gives rise to mental experiences looms as one of the central challenges for science in the new millennium.

It is astounding that cognition and emotion — phenomena that cannot be duplicated in our most sophisticated computers — arise naturally from the electrical activity of large systems of neurons within the brain. Scientific investigation of these phenomena is inherently interdisciplinary, drawing strength from fields as diverse as neurophysiology, cognitive psychology and computational theory. Exciting new findings have emerged in recent decades concerning the neural underpinnings of cognitive functions such as perception, learning, memory, attention, decision-making, language and motor planning, as well as the influence of emotion and motivation upon cognition. With very few exceptions, however, our understanding of these phenomena remains rudimentary. We can identify particular locations in the brain where neuronal activity is modulated in concert with particular external or internal stimuli. In some cases we can even artificially manipulate neural activity in a specific brain structure (using electrical or pharmacological techniques) and cause predictable changes in behaviour. But we encounter substantial difficulties in understanding how modulations in neural activity at one point in the nervous system are actually produced by synaptic interactions between neural systems. Thus our current state of knowledge is somewhat akin to looking out the window of an airplane at night. We can see patches of light from cities and towns scattered across the landscape, we know that roads, railways and telephone wires connect those cities, but we gain little sense of the social, political and economic interactions within and between cities that define a functioning society.

To achieve a more sophisticated level of understanding, investigators must develop new experimental techniques for studying functional interactions between neurons and systems of neurons, and new models for understanding the behaviour of complex, dynamic systems like the brain. Whether major breakthroughs occur on the timescale of years or decades depends substantially on success in developing these new techniques. Irrespective of timescale, an increasingly sophisticated understanding of the neural basis of cognition will influence our society profoundly. It will have practical applications such as treatment of mental disease and



Figure 1 Visual scene of apples on a picnic table. Photograph by David G. Muir.

the design of intelligent machines, and will raise contentious social issues such as the freedom of each individual to choose their behaviour, and the extent to which society can reasonably demand individual responsibility for behaviour.

In what follows, we address three questions. What sorts of cognitive phenomena do neuroscientists seek to explain in terms of brain function? What form do our explanations take, and what technical advances are round the corner to help us? And what are the personal and social implications of understanding the neural underpinnings of consciousness and mental functions?

Visual perception and cognition

The following example of visually based cognition illustrates some of the mental phenomena that neuroscientists seek to understand. Imagine a girl standing before a picnic table (Fig. 1), scrutinizing a jumble of apples in search of the biggest and best granny smith, her favourite kind. She looks, finds one, reaches out and grabs it. The action seems straightforward enough, but this simple description belies the enormous

complexity the brain faces in perceiving the scene, deciding on a course of action, and then executing it.

The brain has no direct access to the rich array of objects and surfaces in the three-dimensional visual environment; it must reconstruct the scene from complex, two-dimensional images falling on the two retinae. Consider just three of the many problems the visual system must solve in scene reconstruction. First, although the retinal images are flat, accurate depth perception is critical for assessing the shapes and sizes of the apples, and for reaching accurately to grab the selected apple. The brain reconstructs the third dimension from multiple cues in the retinal image. For example, the brain can infer depth based on the slight disparity between the images an object casts on the two retinae. Second, in order to pick out the largest apple, the girl must accurately assess the size of each apple. Yet the size of an apple's image on her retina depends on its distance from her. Her visual system takes distance into account and automatically estimates the true physical sizes of objects at different distances, a perceptual phenomenon known as size constancy. Third, the apple scene contains a kaleidoscope of contours: texture elements, colour boundaries, shadows, reflections, specks of dirt, and so on. Only a fraction of these delineate the true boundaries of objects, however, and the brain must sort these out (contour extraction) in order to identify objects accurately.

Once the girl's visual system reconstructs an accurate representation of the scene, a higher-level decision process must evaluate the perceptual information and select one apple in particular. More specifically, her brain must categorize the apples (granny smith, red delicious and so on), assign appropriate affective associations (likes and dislikes) to them, deploy spatial attention to salient objects in the scene, and discriminate fine differences in colour, size and shape to select the best of the granny smiths. Thus, the girl's decision is shaped by immediate sensory information, by previously learned categories drawn from visual memory, and by likes and dislikes based on accumulated experience. Other elements in the scene that are totally irrelevant to the decision must be ignored. Finally, once the brain reconstructs the scene and makes a decision, voluntary

movement systems must plan and execute the appropriate behavioural response (reaching for the selected apple).

Thus, choosing an apple from an apple cart, like innumerable actions we carry out every day, involves a surprising array of cognitive functions. Scientists would like to understand the neural mechanisms that underlie those functions.

Levels of understanding

Understanding the neural basis of a specific cognitive function typically begins with behavioural observations and hypotheses developed by perceptual and cognitive psychologists. Equipped with sound conceptual frameworks originating in behaviour, neurophysiologists can then study underlying brain function at several levels. We will describe three of the most important levels: localization, representation and microcircuitry.

At the coarsest level, the primary issue is 'localization' of function: identifying, for example, neural systems in the brain that are strongly active in response to visual images or when spatial attention is deployed to different regions of a visual scene. As illustrated in Fig. 2a, localization of function has been a dominant theme in brain science, beginning in the nineteenth century when 'phrenologists' attempted to map mental functions onto the brain by correlating aspects of personality and mental ability with the sizes of bumps at different locations on the skull. More reliable evidence for localization of function emerged in the early part of the twentieth century as neurologists learned to recognize mental deficits (sometimes highly specific) that occurred subsequent to damage in particular regions of the brain.

The most commonly used techniques in modern studies of localization are positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), which generate most of the glossy figures in popular science magazines that depict mental activity as coloured blobs on a picture of the brain. Both PET and fMRI measure changes in blood flow to specific regions of the brain while human subjects perform various cognitive tasks¹. The blood flow signal is assumed to reflect changes in metabolic demand resulting from altered levels of neural activity. Using PET and fMRI, investigators can study brain activity in humans at the spatial scale of individual brain structures (a few millimetres) and on a timescale of a few seconds.

The recent avalanche of PET and fMRI papers has produced many insights concerning localization of mental functions (Fig. 2b)²⁻⁴. Although they are substantially more informative than previous findings based on brain damage, distillations of data like the one in Fig. 2b are disconcertingly similar to

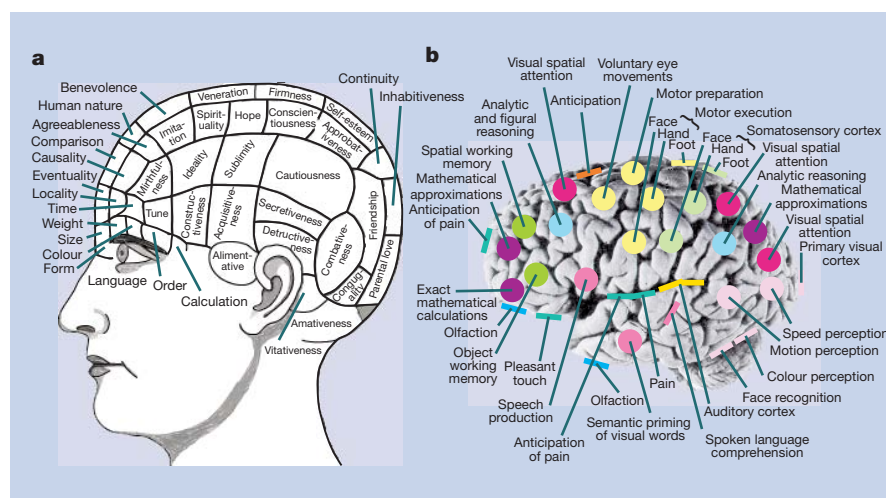


Figure 2 Localization of mental functions. a, According to nineteenth-century phrenologists, lumps on the skull revealed the locations in the brain responsible for mental functions like 'agreeableness' in the frontal lobe and 'cautiousness' in the parietal lobe. b, PET and fMRI studies, which measure changes in cerebral blood flow while various perceptual and cognitive tasks are performed, have revealed a different organization. Imaging data assembled by C. J. Doane.

the one-function-one-brain-area maps of the phrenologists. Will we 'understand' the brain when the map in Fig. 2b is completely filled with blobs? Obviously not; localization data provide little insight into the exact nature of the signals encoded in a given structure, the computations being performed and the interactions between different structures.

Greater insight into how the brain processes and represents information can be obtained by finer-grained studies of individual brain structures using microelectrodes. With these techniques, investigators can measure directly the electrical activity of individual neurons (Fig. 3a), or of small, functionally related clusters of neurons called 'cortical columns'. For example, Fig. 3b schematizes the electrical discharges (black vertical tick marks) of a cortical neuron that responds vigorously to a bar of light sweeping downward across a viewing screen, but not to upward motion of the bar. This neuron, therefore, is 'selective' for downward visual motion. Other neurons are selective for other directions of motion, specific colours, the orientation of line segments, the density of visual texture and many other visual features. Their selectivity allows them to signal or represent the presence or absence of particular features in the visual environment. This 'representational' level of analysis has provided some of the major success stories of systems neuroscience^{5,6}, including descriptions of early stages of visual processing relevant to the problems of contour extraction and depth perception described in the preceding section.

Investigation at the representational level is particularly powerful when carried out in alert animals trained to perform interesting cognitive tasks. For example, researchers

have discovered that certain neurons in the temporal and frontal lobes of the brain are active during short-term memories of specific objects or places⁷⁻⁹, whereas other neurons in the parietal lobe respond dramatically when the animal deploys attention to one or another region of a visual scene¹⁰⁻¹². These functional properties can change strikingly from one cortical column to the next within a given brain area (a spatial scale of several tens of micrometres), revealing an intricate level of organization invisible to fMRI and PET.

These discoveries provide researchers with a deeper 'point of entry' for analysis of the neural systems underlying a particular cognitive function. Transformations in the information carried by single neurons can be inferred by recording from successive brain areas in a particular pathway while the animal performs a task. On the basis of such measurements, investigators can develop quantitative models of how these transformations take place. Furthermore, investigators can test hypotheses about the function of a particular pathway by applying electrical or pharmacological techniques to activate or inactivate discrete clusters of neurons. In our laboratory, for example, we have found that electrically activating clusters of direction-selective neurons through a microelectrode (Fig. 3a) can induce rhesus monkeys to report seeing motion in the direction encoded by the activated neurons¹³. Results of this sort, perhaps more decisively than any others, establish a causal link between the activity of particular classes of neurons and specific mental phenomena.

But even this level of analysis begs fundamental questions about how signals are created, encoded and transmitted by single neurons and assemblies of neurons.

Individual cortical neurons (like the one in Fig. 3a) receive input signals from as many as three to ten thousand other neurons. In a typical experiment, however, the neurophysiologist can characterize the responses of only a few neurons at the tip of the electrode. With such a limited data set, it is difficult to determine exactly how the thousands of synaptic inputs to a cell are transformed to create the cell's pattern of output activity.

This third level of analysis — the detailed 'microcircuitry' and dynamic interactions that give rise to the observed activity of single neurons — is the most vexing for systems and cognitive neuroscience. The challenge stems in part from the extreme complexity of even a single cortical neuron, and in part from daunting practical problems of physical accessibility and visualization in an intact, behaving animal. Elegant intracellular recording techniques can be used to study interactions between neurons in simple preparations (for example, thin slices of tissue removed from the brain, or simple invertebrate nervous systems). The sequence of images in Fig. 3c, for example, shows synaptic events within individual compartments of a single cell in the brain of the common housefly. Each compartment receives synaptic signals from a different input neuron. This neuron is selective for downward visual motion, just like the one in Fig. 3b. In the fly brain, however, it is possible to monitor activity sweeping through the individual compartments of the cell as a bar is swept downward through the visual field. Observations of this kind will shed light on how individual neurons transform the input signals they receive from other neurons into an output signal. Even so, understanding the neural microcircuitry underlying complex cognitive phenomena like decision-making in mammals seems a distant hope, although new technical advances could substantially influence this level of analysis.

A look over the horizon

The immediate impediments to progress in systems and cognitive neuroscience are more technical than conceptual. Computational theorists and psychologists have developed plausible models for many cognitive functions; our primary problem lies in acquiring and analysing the neural data needed to evaluate these models. It will be a sign of real progress if this situation reverses over the next two decades, with conceptual issues coming to the fore.

Technical innovation is needed at all levels of analysis. At the level of localization, improvements in fMRI technology could permit study at the spatial resolution of the cortical column, where so much functional specialization resides. This capability is possible in principle and should emerge gradually over the next decade. More problematic, however, is the temporal resolution

of fMRI, which is limited by blood-flow dynamics to a scale of seconds. Neural processing, in contrast, occurs on a scale of milliseconds. Our best prospects for recovering temporal resolution lie in combining fMRI with techniques that measure electrical activity directly, such as evoked potential recording¹⁴ and magnetoencephalography¹⁵.

At the level of representational analysis, multi-electrode recording techniques are enabling investigators to obtain data simultaneously from larger populations of neurons, and may provide the most promising approach for analysing dynamic interactions between cortical areas.

Optical techniques are likely to become increasingly important both at the representational level of analysis and at the local circuit level. Currently, CCD cameras are capable of gathering reflected (or fluorescent) light from the brain surface, thereby enabling study of brain function at the level of individual cortical columns¹⁶. Study of single neuron dynamics will increasingly rely on multi-photon microscopy, which can isolate signals from individual cells *in vivo* hundreds of microns below the cortical surface¹⁷. Focal stimulation of individual neurons will increasingly involve optical release of caged neurotransmitters¹⁸. The rapid development of new optical probes, caged neuroactive compounds, and light sources and detectors will extend current resolution limits and may permit the application of optical methods to alert animals.

Implications for society

Progress in understanding higher brain functions, fuelled by technological innovation, will certainly exert a major impact on society in the coming century. New discoveries will be important for the diagnosis and treatment of psychiatric and neurological disease, for improving human learning and communication and for informing the design of intelligent machines. But a scientific understanding of the human mind and human behaviour in terms of brain function could also have a profound impact on how we understand ourselves and our society. Two aspects of mentality deserve special consideration: conscious experience and decision-making.

Consciousness

The relationship between conscious experience and brain function is one of the great remaining mysteries of the natural world. How can three pounds of tissue that scientists study with microelectrodes, microscopes and magnetic resonance machines give rise to conscious experience? Indeed, what exactly do we mean by 'conscious experience'?

A classic thought experiment^{19,20} illustrates the problem. Imagine a time in the future when visual neuroscience has come to

completion, when everything is known about how the nervous system responds to light and how it produces visually guided behaviour. Now imagine a colour-blind neuroscientist in this golden era who can give a complete account of the neural events underlying a normal observer's ability to discriminate and identify colours. Now suppose our scientist's colour-blindness were miraculously cured and he saw green for the first time. The question is, would he learn anything new about colour perception? Most would answer 'yes': what he learns is precisely the conscious experience of colour. Consciousness is the aspect of our mental life that we can only understand through subjective first-person experience.

Can we study consciousness scientifically? This is controversial among researchers, but we believe that progress will occur only if investigators can develop operational

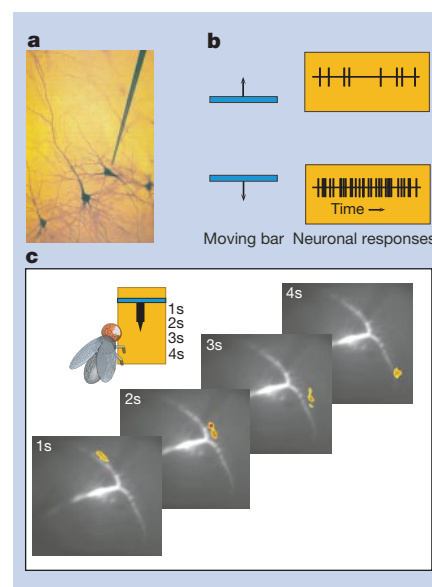


Figure 3 Illustration of the representational and microcircuit levels of analysis. **a**, Pyramidal neurons in monkey visual cortex, stained by the Golgi method. A tungsten microelectrode typically used for extracellular recording has been superimposed (image from ref. 5). **b**, Schematic of a bar stimulus (blue bar) moving either up or down and the response of a direction-selective neuron. Black tick marks represent the electrical discharges of the neuron. The neuron is direction selective, discharging more vigorously for downward motion than for upward motion. **c**, Tangential cell in a housefly in response to a bar moving downward in the visual field (the preferred direction for that cell). The cell was filled with calcium green, which fluoresces in response to influxes of calcium caused by synaptic inputs from other neurons. The four images represent four successive 1-second intervals as the bar sweeps through the visual field producing electrical changes in different spatial compartments of the cell (S. Single and A. Borst, unpublished data).

definitions of conscious experience, as has been accomplished for other aspects of cognition. In a classical form of learning, for example, Pavlov's dog generated a measurable dependent behaviour (salivation) in response to controlled manipulation of independent variables (pairing of bell with food). In general, we can hope to understand the neural basis of any cognitive phenomenon that can be defined operationally in a similar manner.

Accomplishing this for consciousness will be difficult, because conscious experience is intrinsically subjective. A person's verbal reports of her conscious experiences may be sufficiently reliable to serve as objective observations for testing hypotheses about consciousness. But the notion that verbal reports (or any other overt behaviour) are an adequate proxy for conscious experiences is itself not scientifically testable. We must therefore acknowledge that this assumption is extra-scientific (as is parsimony, another favourite assumption of scientists). Even if we do accept this assumption in general, verbal reports may fall short under some circumstances. For example, if one person reacts to pain more readily than another, is it because he has a more intense sensation of pain, or because his threshold for reacting to pain is lower? Could any verbal exchange answer such questions? Although animals can be trained to perform simple behavioural tasks, these sorts of questions only become more perplexing without the benefit of language. Therefore, any measure of consciousness in animals is likely to be controversial, despite the common intuition that the family dog enjoys conscious experiences of one sort or another.

Ultimately, no matter how precisely we manage to link verbal reports of conscious experience to brain activity, fundamental mysteries are likely to persist²¹. Exactly how something as ineffable as subjective consciousness can arise from macromolecules, synapses and action potentials may well remain a conundrum.

Decision-making and determinism

An exciting frontier in the study of higher brain function is the attempt to understand mechanistically how decisions are formed. 'Decision processes' are the key cognitive links between perception and action. Perceptual processes carry out the functions of scene reconstruction, contour extraction, and so on, and motor processes implement the planning and execution of a behavioural response. Intermediate levels of the system, however, must evaluate the sensory evidence represented in early cortical areas, and 'decide', for example, which apple is to be the target of a reaching movement.

Neurophysiologists have now begun to investigate the neural underpinnings of decision processes in monkeys trained to

perform simple discrimination tasks. By recording at successive stages of known anatomical pathways linking sensory to motor areas of the cortex, investigators have uncovered intriguing evidence for neural systems that represent what the monkey 'decides' about the stimulus as opposed to what the stimulus actually is²²⁻²⁴. The neural activity recorded in such experiments can actually predict the outcome of the monkey's decision several seconds before it is revealed by a behavioural response. In addition, neurophysiologists are now taking notice of a fact that psychologists and economists have long recognized: decisions are informed as much by reward expectation and personal history as by the sensory evidence available at any given time. If, for example, an animal knows from past experience that choice A tends to be rewarded more frequently than choice B, it may choose A even if countervailing sensory evidence suggests that choice B will be rewarded on the current trial. Important new discoveries have been made concerning neural signals related to reward expectation and relative reward potency²⁵⁻²⁷; indeed such signals have been found in the inferior parietal cortex of monkeys²⁸, a cortical region implicated in perceptual decision-making.

Limits to determinism?

The next decade will certainly witness substantial progress in understanding how decisions are formed within the brain, but these studies in particular raise disquieting questions. The crux of the problem lies in the implications of physical determinism for our concepts of personal freedom and moral responsibility. If the most sophisticated aspects of our mental lives, from decisions as trivial as selecting an apple to those as important as choosing a spouse, are determined by the molecular and cellular events that generate electrical activity in the nervous system, what are we to make of our subjective experience of freedom? If our sense of freedom in making moral decisions is illusory, can anyone reasonably be held responsible for his or her actions? Can a criminal reasonably be punished for actions over which he has no meaningful control?

We have no solution to this problem, but several perspectives seem important to bear in mind. The most significant, perhaps, is that both points of view — physical determinism and personal freedom — are utterly essential in humanity's attempt to understand itself. The assumption of cause-and-effect determinism is fundamental not only to science but also to everyday life. For example, every day most of us must negotiate streets full of busy traffic in order to survive until the next day. The basic perceptual, decision-making and motor mechanisms within our brains must be sufficiently deterministic that we take the appropriate action every time. Furthermore, the assumption of

determinism underlies our hope of finding cures for devastating psychiatric and neurological diseases. If these conditions do not have physical causes within the brain, there is no reason to hope that they can be cured by physical (that is, medical) interventions.

On the other hand, the assumption of a meaningful degree of personal freedom is essential not only to our personal and social lives, but to science as well. Scientists require the freedom to evaluate data and to reject false hypotheses. But if our mental processes unfold with the physical determinism of a machine, what guarantee do we have that the beliefs the machine generates, by the scientific method or otherwise, are true? And if the machine generates false beliefs (the belief in determinism, perhaps?) how would we discard them? We are left, then, with the paradox that both perspectives seem necessary, for the community of science as well as for everyday life. Perhaps these perspectives will be reconciled at some point in the future. For the present, our only obvious option is to live with both, and accept the paradox as a leavening dose of humility in our intellectual lives.

M. James Nichols and William T. Newsome are at the Howard Hughes Medical Institute and Department of Neurobiology, Stanford University School of Medicine, Stanford, California 94305, USA.

e-mail: bill@monkeybiz.stanford.edu

- Ogawa, S. *et al. Proc. Natl Acad. Sci. USA* **89**, 5951-5955 (1992).
- Posner, M. I. & Raichle, M. E. *Proc. Natl Acad. Sci. USA* **95**, 763-764 (1998).
- Sell, L. A. *et al. Eur. J. Neurosci.* **11**, 1042-1048 (1999).
- Wandell, B. A. *Annu. Rev. Neurosci.* **22**, 145-173 (1999).
- Hubel, D. H. *Eye, Brain, and Vision* (Scientific American Library, New York, 1995).
- Parker, A. J. & Newsome, W. T. *Annu. Rev. Neurosci.* **21**, 227-277 (1998).
- Tanaka, K. *Curr. Opin. Neurobiol.* **2**, 502-505 (1992).
- Fuster, J. M. *Memory in the Cerebral Cortex* (MIT Press, Cambridge, 1995).
- Goldman-Rakic, P. S. *Proc. Natl Acad. Sci. USA* **93**, 13473-13480 (1996).
- Desimone, R. & Duncan, J. *Annu. Rev. Neurosci.* **18**, 193-222 (1995).
- Maunsell, J. H. *Science* **270**, 764-769 (1995).
- Colby, C. L. & Goldberg, M. E. *Annu. Rev. Neurosci.* **22**, 319-349 (1999).
- Salzman, C. D., Murasugi, C. M., Britten, K. H. & Newsome, W. T. *J. Neurosci.* **12**, 2331-2355 (1992).
- Martinez *et al. Nature Neurosci.* **2**, 364-369 (1999).
- George, J. S. *et al. J. Clin. Neurophysiol.* **12**, 406-431 (1995).
- Hubener, M., Shoham, D., Grinvald, A. & Bonhoeffer, T. *J. Neurosci.* **17**, 9270-9284 (1997).
- Svoboda, K., Helmchen, F., Denk, W. & Tank, D. W. *Nature Neurosci.* **2**, 65-73 (1999).
- Wang, S. S. & Augustine, G. J. *Neuron* **15**, 755-760 (1995).
- Jackson, F. *Phil. Quarterly* **32**, 127-136 (1982).
- Thompson, E. *Phil. Studies* **68**, 321-349 (1992).
- Nagel, T. *Phil. Rev.* **4**, 435-450 (1974).
- Shadlen, M. N. & Newsome, W. T. *Proc. Natl Acad. Sci. USA* **93**, 628-633 (1996).
- Romo, R. & Salinas, E. *Curr. Opin. Neurobiol.* **9**, 487-493 (1999).
- Schall, J. D. & Thompson, K. G. *Annu. Rev. Neurosci.* **22**, 241-260 (1999).
- Gallistel, C. R. *Cognition* **50**, 151-170 (1994).
- Schultz, W., Dayan, P. & Montague, P. R. *Science* **275**, 1593-1599 (1997).
- Schultz, W. *J. Neurophysiol.* **80**, 1-27 (1998).
- Platt, M. L. & Glimcher, P. W. *Nature* **400**, 233-238 (1999).

Acknowledgements. We thank M. Schnitzer for useful suggestions during the preparation of the manuscript.

'Earth system' analysis and the second Copernican revolution

H. J. Schellnhuber

Optical magnification instruments once brought about the Copernican revolution that put the Earth in its correct astrophysical context. Sophisticated information-compression techniques including simulation modelling are now ushering in a second 'Copernican' revolution. The latter strives to understand the 'Earth system' as a whole and to develop, on this cognitive basis, concepts for global environmental management.

There are many ways of looking forward in time. One of the most amusing (and sometimes terrifying) is the 'forward-view mirror' — contemplation of the future by reflecting on the past. If we consider the unravelling of the mysteries of the human body by physicians over the past three millennia, we see much that is relevant to unravelling the mysteries of the Earth's physique, or 'Gaia's body'¹.

With our present-day understanding of medical science, it seems incredible that the Hippocratic school, which based analysis and prognostics of the human body on the 'composition of the humours' of individual patients, dominated Western medicine well beyond the Renaissance. Great leaps in knowledge, such as Vesalius's anatomical revelations published in 1542, or Harvey's physiological studies in 1628, were ignored or suppressed — notably by the deans of Paris's 'infallible' medical faculty. The first scientific treatise relating contagious diseases to activities of microorganisms, rather than harmful vapours, or 'miasmas', did not appear until 1840².

When the Enlightenment came, its ultimate triumph was based, literally, on light — the ability to process radiation received from objects of specific interest. The invention of the operational microscope in 1608 by the Dutch spectacle-maker Zacharias Jansen, realizing a proposal made in 1267 by Roger Bacon, was a turning-point in scientific history. For the first time the human eye could transcend its natural limits and begin to explore the wonders of the microcosmos.

Many more wonders awaited revelation through the processing of light — above all, the spangled nocturnal heavens with their billions of gigantic entities made so tiny by distance. Once again, faint rays of light emitted by objects had to be invigorated through ingenious devices, in this case telescopes. And so optical amplification techniques brought about the great Copernican revolution, which finally put the

Earth in its correct astrophysical context.

Today, some 500 years after Copernicus, Cusanus and company, our civilization sets

about visiting neighbouring planets, scrutinizing stellar objects at the brink of creation, and even tracking down extraterrestrial

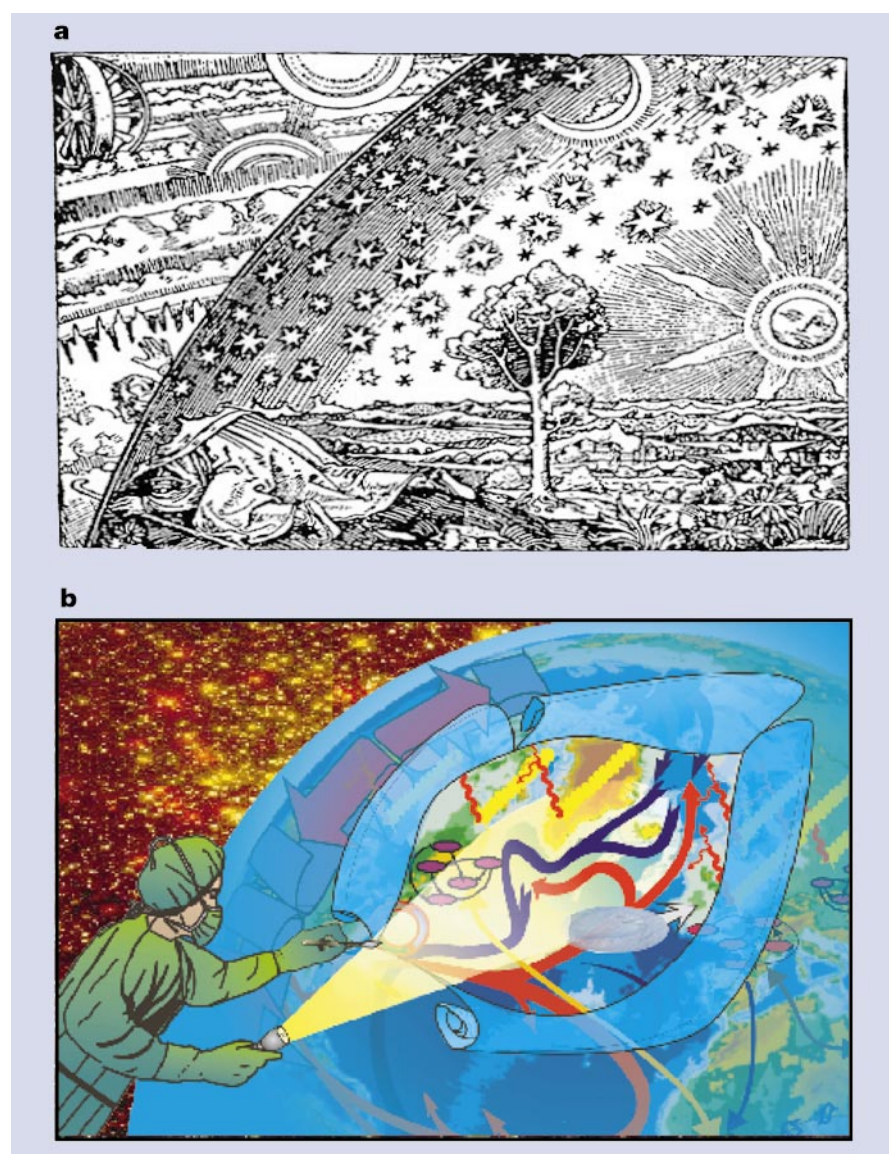


Figure 1 A tale of two revolutions. a, The shock of the Enlightenment as expressed in a (probably apocryphal²⁶) fifteenth-century woodcut. b, 'Earth-system' diagnostics in the twenty-first century.

intelligence. This spectacular augmentation of the Renaissance impulse is being accompanied by a crescendo of other scientific activities which will soon culminate in a second 'Copernican' revolution. This new revolution will be in a way a reversal of the first: it will enable us to look back on our planet to perceive one single, complex, dissipative, dynamic entity, far from thermodynamic equilibrium — the 'Earth system'. It may well be nature's sole successful attempt at building a robust geosphere-biosphere complex (the ecosphere) in our Galaxy, topped by a life-form that is appropriately tailored for explaining the existence of that complex, and of itself.

Such an explanation needs eventually to encompass all the pertinent processes linking the system's constituents at all scales — from convection deep in the Earth's mantle to fluctuations at the outer limit of the atmosphere. New instruments are necessary, especially macroscopes³ which reduce, rather than magnify as microscopes do, giving Earth-system scientists an objective distance from their specimens — no longer too close for cognitive comfort.

There are three distinct ways to achieve 'holistic' perceptions of the planetary inventory, including human civilization:

1. The 'bird's-eye' principle.

An obvious trick for obtaining a panoramic view of the Earth is to leave it, and observe it from a distance. The race to the Moon in the 1960s opened up the opportunities for this particular macroscope technique, and created the now-familiar image of our blue planet floating in the middle of a dark, cold nowhere. The trick was not exactly cheap, however; NASA's lunar ventures absorbed some US\$95 billion overall. Now, space stations, shuttles and an armada of smart

satellites are about to establish the details of a complete Earth reconnaissance.

2. The digital-mimicry principle.

A more sophisticated, and less expensive, macroscope technique is simulation modelling. Here, components and processes of the original Earth system are replaced by mathematical representatives as accurate as our evolving knowledge allows. These formal chimaeras are then animated electronically, to imitate the dynamic complex of real relationships, in virtual space-time. The menagerie of Earth-system models includes tutorial, conceptual and 'analogical' specimens⁴. One significant advantage of this macroscope is that it allows a multitude of potential planetary futures to be played out, with no more a risk than a computer crash. The validation of Earth-simulation machines remains problematic, although relentless training with palaeorecords can teach them satisfactory hind-casting skills (see, for example, refs 5–7).

3. The 'Lilliput' principle.

As a third option, there is the 'incredible shrinking Earth' idea, as enacted in the Sonora Desert in Biosphere II (ref. 8). The idea involves rebuilding the ecosphere in flesh, blood and rock, on a scale reduced by many orders of magnitude. Such a nano-planet can be conveniently scrutinized for operational stability or emerging self-organization processes. Despite its disastrous performance, the Biosphere-II project provoked fresh scientific attitudes towards life. And in fact, the free-air experiments⁹, which are currently being used to investigate the effect of atmospheric CO₂ enrichment on agroecosystems and forests, subscribe to a similar empiricist philosophy.

Most probably, the future of macroscopes

will be dominated by intelligent combinations of these principles, particularly the first two. Planetary monitoring — by remote sensing and a worldwide net of *in situ* measurement devices — will be complemented and synchronized by data models to generate a continuously updated digital 'Weltbild'.

The quasi-antithetical spirits of the first and second Copernican revolutions may be visualized by contrasting a famous ancient allegory with a modern cartoon (Fig. 1). The explorer featured in Fig. 1b is dressed as a doctor for two reasons. First, the continuing investigation into the Earth's physique is in many respects reminiscent of the exploration of the human body during the Renaissance. Science historians looking back from, say, AD 2300, will tell yet again a tale of incredible delusions and triumphs. And second, a significant impetus behind the second Copernican revolution is the insight that the ecosphere's operation may be being transformed qualitatively by human interference. So the macroscope is a diagnostic instrument, generating evidence necessary for treatment¹⁰.

This means that we are confronted ultimately with a control problem, a geo-cybernetic task that can be summed up in three fundamental questions¹¹. First, what kind of world do we have? Second, what kind of world do we want? Third, what must we do to get there?

These questions indicate the immensity of the challenge posed by Earth-system analysis¹². I would like now to narrow our gaze to a few crucial aspects of this transdisciplinary adventure, using the light of the latest progress in pertinent science, particularly that orchestrated by the International Geosphere-Biosphere Programme (IGBP; see Box 1)¹³.

Box 1 The International Geosphere-Biosphere Programme (IGBP)

Each day seems to bring new discoveries — of mega-fluctuations, long-distance teleconnections, feedback loops or phase-transition lines in the entrails of the planetary ecosystem. This development is driven predominantly by IGBP research and is shedding new light on past, present and future global changes. For example, one IGBP core project involves reconstructing, in great detail, global and regional climatic history as far back as 500,000 years ago²⁷. Another major project is more concerned with the current operational mode of the Earth system. It tries to solve the 'puzzle' of oceanic CO₂ flux in a geographically explicit way. Intermediate results indicate that CO₂ is upwelling and leaving the ocean in the subarctic western Pacific Ocean in winter, in the Persian Gulf in summer, and west of South America all year round. In contrast, oceanic regions, where warm currents like the Gulf Stream and the Kuroshio are being cooled, take up large amounts of CO₂ (ref. 28).

Complementary activities scrutinize the CO₂ metabolism of the terrestrial biosphere. Recent findings provide hints that the continental ecosystems may turn into a carbon source as a result of climate change at some time in the next century. A subtle combination of factors, including changes in fire, storms and pest infestations, is responsible for this disconcerting prospect²⁹. Another disturbing result was obtained in the coastal zone: model-supported forecasting of the geochemical effects on coral reefs of anthropogenic CO₂ build-up in the atmosphere demonstrates the high vulnerability of these ecosystems. Crucial mechanisms for reef accumulation in the tropics could be weakened by 30 per cent by 2050³⁰.

Understanding the Earth system

At the highest level of abstraction, the make-up of the Earth system *E* can be represented by the following 'equation':

$$E = (N, H) \quad (1)$$

where $N = (a, b, c, \dots)$; $H = (A, S)$.

This formula expresses the elementary insight that the overall system contains two main components, namely the ecosphere *N* and the human factor *H*. *N* consists of an alphabet of intricately linked planetary sub-spheres *a* (atmosphere), *b* (biosphere), *c* (cryosphere; that is, all the frozen water of Earth), and so on. The human factor is even more subtle: *H* embraces the 'physical' sub-component *A* ('anthroposphere' as the aggregate of all individual human lives, actions and products) and the 'metaphysical' sub-component *S* reflecting the emergence of a 'global subject'. This subject manifests itself, for instance, by



Although effects such as the glaciations may still be interpreted as over-reactions to small disturbances — a kind of cathartic geophysiological fever — the main events, resulting in accelerated maturation by shock

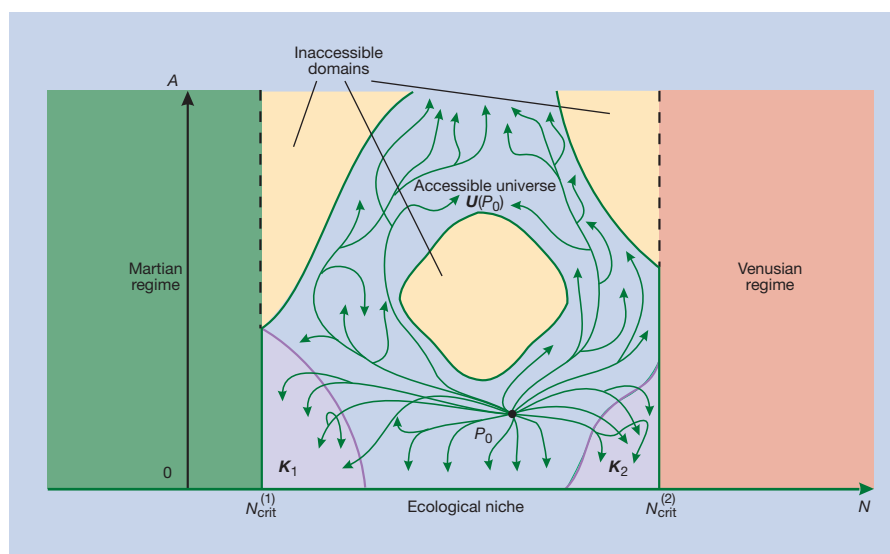


Figure 3 A 'theatre world' for representing paradigms of sustainable development. The space of all conceivable co-evolution states $P = (N, A)$ is spanned by a 'natural' axis, N (representing, say, global mean temperature) and a civilizational axis, A (representing, say, global gross product). Vertical lines at $N_{crit}^{(1)}$ and $N_{crit}^{(2)}$ delimit the niche of subsistence states for humanity between an ultra-cold 'Martian regime' and an ultra-hot 'Venusian regime'. The domain $U(P_0)$ ('accessible universe') embraces all possible co-evolution states that can be reached from the present state P_0 by some management sequence from the overall pool. $U(P_0)$ contains specific 'catastrophe domains' K_1 and K_2 .

treatment, indicate that Gaia faces a powerful antagonist. Rampino has proposed personifying this opposition as Shiva, the Hindu god of destruction¹⁹.

About four billion years into Earth's history, a third planetary might emerged, a challenger to these two intransigent forces: human civilization. Let us stay with mythological imagery and call this power Prometheus. For the purposes of Earth-system analysis, Prometheus is to be identified with the mega-factor H in equation (1).

There is no need to reel off a tally of environmental folklore about the influence of Prometheus on the ecosphere: the ozone lesson is enough to show that humanity is indeed capable of modifying N at a strategic level. There is overwhelming evidence that the stratospheric ozone shield against harmful UV-B radiation has been perforated accidentally by industrial by-products, in particular by chlorofluorocarbons (CFCs). The physicochemical processes causing this disturbing effect are intricate; they were deciphered only a few years ago and the elucidators awarded a Nobel prize. But "things could be much worse", as one of the laureates, Paul Crutzen, puts it: "Bromine is almost a hundred times more dangerous for ozone than chlorine. If the chemical industry had developed organobromine compounds instead of CFCs then...we would have been faced with a catastrophic ozone hole everywhere and all year-round during the 1970s — probably before atmospheric scientists had developed the knowledge necessary to identify the problem."²⁰

We have been extremely lucky in being

granted the chance of correcting our ways through the Montreal ozone-protection protocol. Ironically, the eminent meteorologist Chapman suggested back in 1934 creating artificial ozone holes by injecting appropriate ozone depleters into the stratosphere as a way of improving astronomy's ultraviolet remote-sensing accuracy²¹.

Global environmental change is all around us now, and the material components of the Earth system, N and A , are behaving like a strongly coupled complex. To assess the crucial consequences of this phase transition in planetary history, we must, and can, do better than drawing wiring diagrams of the Bretherton type. Quite recently, IGBP launched a promising macro-scope-making initiative aimed at advancing Earth-system models of intermediate complexity (EMICs; see Box 2). These models seek to integrate the main processes and forces — Gaia, Shiva and Prometheus — through effective quantitative equations. Ideally, they would simulate all of the pertinent features of the N - A complex on the system's level, and operate fast enough to serve as 'time machines' — producing virtual vistas of the far environmental past and future^{6,7}.

It has become clear, however, that our quantitative mimicry skills for anthroposphere processes (like industrial growth or transmigration dynamics) lag far behind our ecosphere-simulation capacity. The most capable sub-models currently available for representing relevant socioeconomic components of the Earth system are those used in 'integrated assessments of climate

change'²². But these instruments are still far too crude — a problem scientific progress must soon remedy. Massive intellectual and financial investments should generate adequate (long-term, non-equilibrium, multi-actor) modules within the next two decades.

Controlling the Earth system

Now assume that the research community does its job and develops a perfect hierarchy of transdisciplinary EMICs. Who will use the insights, the hindsight and foresights generated by these time machines? And in what way? These questions prompt us to revisit the human factor H in the Earth-system equation (1). Formally, $H = (A, S)$, where A reflects the physiological-metabolic contribution of global civilization to planetary operation. This contribution is qualitatively not dissimilar to the role played by the sheep of the world, which reflect sunlight (albedo effect), overgraze pastures (soil degradation) and emit the powerful greenhouse gas methane.

But H embraces a second sub-factor, S , which makes all the difference. This entity, introduced as the 'global subject' above, represents the collective action of humanity as a self-conscious control force that has conquered our planet. The global subject is real, although immaterial. One key to its emergence from the physical basis is world-wide communication. As you read this essay, you are engaging in an indirect dialogue with the author. That is insignificant, however, compared with the direct global 'polylogue' taking place via the Internet. Global telecommunication will ultimately establish a cooperative system generating values, preferences and decisions as crucial commonalities of humanity online.

The building and application of macroscopes will be of tremendous help to the global subject in finding its identity. An ever-evolving Earth-observation system²³ will allow S to watch its own footprints on the ecosphere, and Earth-simulation models will enable S to make collective 'rational choices' on the system's level. Finally, densely linked global institutions, as well as innumerable worldwide activists' networks, will help enforce resolutions of S , such as those made in international environmental conventions. This is the emergence of a modern 'Leviathan', embodying teledemocracy³ and putting the seventeenth-century imagination of the English philosopher Thomas Hobbes into the shade.

The global subject will reign over the centuries to come. One of its most responsible tasks will be to seek out a tolerable environmental future from the infinity of optional co-evolutions of N and A . In other words, S must guarantee sustainable development. In spite of, or because of, its fuzziness, the notion of sustainable

Box 2 Intermediate-complexity modelling

When devising Earth-system models, scientists have to resist two fatal attractions. The first one is over-simplification, which tends to ignore even crucial elements of planetary dynamics like the episodic warming of the tropical East Pacific ('El Niño'; see ref. 31). The second one is over-sophistication, often resulting in bulky and expensive models that defy a transparent analysis of complexity by averaging over the details irrelevant to the issue studied. This can be achieved by constructing reduced-form representations of the full set of dynamic equations according to well known principles from scientific computing. As far as the atmosphere module is concerned there exists, for instance, the option to separate 'slow' from 'fast' variables³² or to apply filtering techniques to the primitive equations of motion³³.

development has become an end-of-millennium *idée fixe*. Entire libraries could be crammed with treatises devoted to the topic.

Dissatisfied with the lack of systematics in the overall sustainability debate, I embarked, a couple of years ago, on the quest of developing a rigorous common formalism, extracting the essence of all possible concepts. On the basis of such a formalism, the elaboration of 'mathematical sustainable-development ethics' becomes feasible^{4,12}.

Some central results can be illustrated in a two-dimensional 'theatre world' where sustainable development is played out as a strategic planning exercise (Fig. 3). Two important aspects should be emphasized. First, the overall co-evolution space includes several unpleasant domains: apocalyptic zones as exemplified by the Martian regime (hypothetically attainable through a runaway cooling-chamber process), or the Venusian regime (hypothetically attainable through a runaway greenhouse process)¹⁶, and catastrophe zones, where humanity might subsist, but in a miserable manner. Second, there exists a non-trivial subspace, the 'accessible universe', which consists of the ensemble of all realizable future co-evolution paths. These are generated by distinct management options contained in the strategic pool at the disposal of the global subject. Note that 'business as usual', that is, planless ecosphere transformation through individual opportunism, is definitely one of these options.

Five competing paradigms for sustainable development can be identified. (1) Standardization — prescribing an explicit long-term co-evolution corridor emanating from P_0 in Fig. 3. (2) Optimization — getting the best, that is, maximizing an aggregated

anthroposphere–ecosphere welfare function by choosing the proper co-evolution segment over a fixed time period. (3) Pessimization — avoiding the worst, that is, steering well clear of catastrophe domains, allowing for the possibility of bad management in the future. (4) Equitization — preserving the options for future generations, not contracting the 'accessible universe' over time. (5) Stabilization — bringing the $N-A$ complex into a desirable state in co-evolution space and maintaining it there by good management.

So, here is the menu from which humanity can select its master principle, or suitable combinations thereof, for Earth-system control. The formal elaboration of these paradigms and putting them into operation using, for example, EMICs, is a highly non-trivial exercise. And there still remains the question "What must we do?". It would be foolish to try to answer this comprehensively, but a few eclectic suggestions may give an idea of the scope of the challenge. I give these in descending order of speculativeness:

Optimization.

The present geostrategic design of the $N-A$ complex is not optimal. The industrialized countries dominate the fields of agricultural production, technological innovation, environmental protection, tourist services and many other areas. This implies not only considerable inequity across the Earth's population, but also unstable distortions of the natural web of energetic and material fluxes. A global redesign could aim at establishing a more 'organic' distribution of labour, where the temperate countries are the main producers of global food supplies, the sub-tropical zones produce renewable energies and high technology, and the tropical zones preserve biodiversity and offer recreation.

Stabilization.

Why should Prometheus not hasten to Gaia's assistance? Geoengineering proposals have become popular as a way of mitigating the anthropogenic aberrations of the ecosphere. One interesting idea features iron fertilization of certain ocean regions to stimulate the marine biological pump which draws down CO_2 (ref. 24). And Russian scientists are currently elaborating a repair scheme for the ozone layer using orbital lasers. But we can also think of proactive control of natural planetary variability: insights acquired during the present climate crisis may enable humanity to suppress future glaciation events by judicious injection of 'designer greenhouse gases' into the atmosphere.

Pessimization.

Least speculative and most essential is the creation of a manual of minimum safety standards for operating the Earth system. Human interference with the ecosphere may provoke the perhaps irreversible transgression of critical thresholds, bringing about

qualitatively different environmental conditions on a large scale. The Intergovernmental Panel on Climate Change is now addressing risks like the shut-down of the ocean conveyor belt, the destabilization of the West Antarctic ice sheet, and the abrupt die-back of tropical forests due to super-critical global warming. The results of IGBP and related global research programmes will soon enable us to identify and respect 'guardrails'²⁵ for responsible planetary management.

In fact, the second Copernican revolution will be completed only if we take this responsibility — in spite of irreducible cognitive deficits as once lamented by Alonso X of Castile: "If the Lord Almighty had consulted me before embarking on the Creation, I would have recommended something simpler"¹⁴.

H. J. Schellnhuber is at the Potsdam Institute for Climate Impact Research, Telegrafenberg C4, PO Box 60 12 03, D-14412 Potsdam, Germany.

e-mail: john@pik-potsdam.de

1. Volk, T. *Gaia's body* (Copernicus, New York, 1998).
2. Winkle, S. *Geißeln der Menschheit* (Artemis & Winkler, Düsseldorf, 1997).
3. de Rosnay, J. *Le Macroscopie* (Seuil, Paris, 1975); *L'homme Symbiotique* (Seuil, Paris, 1995).
4. Schellnhuber, H. J. & Kropp, J. *Naturwissenschaften* **85**, 411–425 (1998).
5. Prentice, I. C. & Webb, T. J. *Biogeogr.* **25**, 997–1005 (1998).
6. Ganopolski, A., Rahmstorf, S., Petoukhov, V. & Claussen, M. *Nature* **391**, 351–353 (1998).
7. Claussen, M. et al. *Geophys. Res. Lett.* **26**, 2037–2040 (1999).
8. Beardsley, T. *Sci. Am.* **273**(2), 18–20 (1995).
9. Long, S. (ed.) *Plant Cell Environ.* **22** (6; spec. issue 567–755) (1999).
10. Clark, W. C. *Sci. Am.* **261**(3), 18–26 (1989).
11. Blackburn, C. (ed.) *Global Change and the Human Prospect* (Sigma Xi, Research Triangle Park, 1992).
12. Schellnhuber, H. J. & Wenzel, V. (eds) *Earth System Analysis. Integrating Science for Sustainability* (Springer, Berlin, 1998).
13. Newton, P. *Nature* **400**, 399 (1999).
14. Fisher, A. *Mosaic* **19**, 52–59 (1988).
15. Lenton, T. *Nature* **394**, 439–447 (1998).
16. Franck, S. et al. *Tellus B* (in the press).
17. Broecker, W. S. & Peng, T. H. *Greenhouse Puzzles* (Eldigio, Palisades, 1998).
18. Lewis, J. S. *Rain of Iron and Ice* (Addison-Wesley, Reading, 1995).
19. Rampino, M. R. in *Scientists on Gaia* (eds Schneider, S. H. & Boston, P. J.) 382–390 (MIT Press, Cambridge, Massachusetts, 1993).
20. Crutzen, P. J. *Angew. Chem. Int. Ed. Engl.* **35**, 1758–1777 (1996).
21. Chapman, S. Q. J. *R. Meteorol. Soc.* **60**, 127–142 (1934).
22. Schneider, S. H. *Environ. Mod. Assess.* **2**, 229–249 (1997).
23. http://www.gos.udel.edu/publications/select_pub.htm
24. Coale, K. H. et al. *Nature* **383**, 495–501 (1996).
25. Petschel-Held, G., Schellnhuber, H. J., Bruckner, T., Toth, F. L. & Hasselmann, K. *Clim. Change* **41**, 303–331 (1999).
26. Robin, H. *The Scientific Image from Cave to Computer* (Abrams, New York, 1992).
27. Petit, J. R. et al. *Nature* **399**, 429–436 (1999).
28. Takahashi, T. et al. *Proc. Natl Acad. Sci. USA* **94**, 8292–8299 (1997).
29. Walker, B., Steffen, W., Canadell, J. & Ingram, J. *The Terrestrial Biosphere and Global Change* (Cambridge Univ. Press, 1999).
30. Kleypas, J. A. et al. *Science* **284**, 118–120 (1999).
31. Timmermann, A. et al. *Nature* **398**, 694–697 (1999).
32. Saltzman, B. in *Physically-Based Modelling and Simulation of Climate and Climatic Change - Part II* (ed. Schlesinger, M. E.) 737–754 (Kluwer Academic Publishers, 1988).
33. Petoukhov, V. et al. *Clim. Dynamics* (in the press).

Acknowledgements. I thank A. Block, M. Cassel, R. B. Cathcart, W. Cramer, H. Hoff, S. Lütkeemeier, S. Rahmstorf, D. Smart, M. Stock, J. v. Braun, V. Wenzel and the entire IGBP community for intellectual and technical support.

From molecular to modular cell biology

Leland H. Hartwell, John J. Hopfield, Stanislas Leibler and Andrew W. Murray

Cellular functions, such as signal transmission, are carried out by 'modules' made up of many species of interacting molecules. Understanding how modules work has depended on combining phenomenological analysis with molecular studies. General principles that govern the structure and behaviour of modules may be discovered with help from synthetic sciences such as engineering and computer science, from stronger interactions between experiment and theory in cell biology, and from an appreciation of evolutionary constraints.

Although living systems obey the laws of physics and chemistry, the notion of function or purpose differentiates biology from other natural sciences. Organisms exist to reproduce, whereas, outside religious belief, rocks and stars have no purpose. Selection for function has produced the living cell, with a unique set of properties that distinguish it from inanimate systems of interacting molecules. Cells exist far from thermal equilibrium by harvesting energy from their environment. They are composed of thousands of different types of molecule. They contain information for their survival and reproduction, in the form of their DNA. Their interactions with the environment depend in a byzantine fashion on this information, and the information and the machinery that interprets it are replicated by reproducing the cell. How do these properties emerge from the interactions between the molecules that make up cells and how are they shaped by evolutionary competition with other cells?

Much of twentieth-century biology has been an attempt to reduce biological phenomena to the behaviour of molecules. This approach is particularly clear in genetics, which began as an investigation into the inheritance of variation, such as differences in the colour of pea seeds and fly eyes. From these studies, geneticists inferred the existence of genes and many of their properties, such as their linear arrangement along the length of a chromosome. Further analysis led to the principles that each gene controls the synthesis of one protein, that DNA contains genetic information, and that the genetic code links the sequence of DNA to the structure of proteins.

Despite the enormous success of this approach, a discrete biological function can only rarely be attributed to an individual molecule, in the sense that the main purpose of haemoglobin is to transport gas molecules in the bloodstream. In contrast, most biological functions arise from interactions among

many components. For example, in the signal transduction system in yeast that converts the detection of a pheromone into the act of mating, there is no single protein responsible for amplifying the input signal provided by the pheromone molecule.

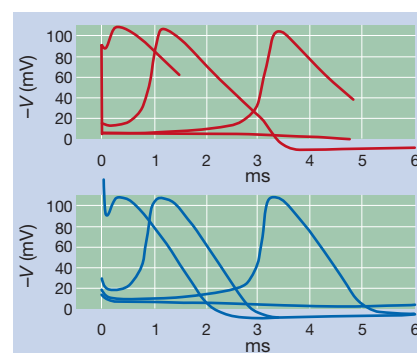
To describe biological functions, we need a vocabulary that contains concepts such as amplification, adaptation, robustness, insulation, error correction and coincidence detection. For example, to decipher how the binding of a few molecules of an attractant to receptors on the surface of a bacterium can make the bacterium move towards the attractant (chemotaxis) will require understanding how cells robustly detect and amplify signals in a noisy environment.

Having described such concepts, we need to explain how they arise from interactions among components in the cell.

We argue here for the recognition of functional 'modules' as a critical level of biological organization. Modules are composed of many types of molecule. They have discrete functions that arise from interactions among their components (proteins, DNA, RNA and small molecules), but these functions cannot easily be predicted by studying the properties of the isolated components. We believe that general 'design principles' — profoundly shaped by the constraints of evolution — govern the structure and function of modules. Finally, the notion of function and functional properties separates biology

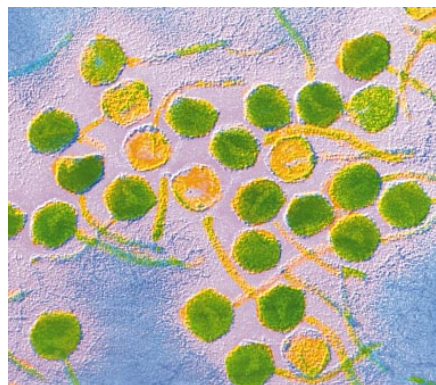
Box 1 Phenomenological analysis of action potentials in nerve cells

Action potentials are large, brief, highly nonlinear pulses of cell electrical potential which are central to communication between nerve cells. Hodgkin and Huxley's analysis of action potentials²⁹ exemplifies understanding through *in silico* reconstruction. They studied the dynamical behaviour of the voltage-dependent conductivity of a nerve cell membrane for Na⁺ and K⁺ ions, and described this behaviour in a set of empirically based equations. At the time, there was no information available about the channel proteins in nerve cell membranes that are now known to cause these dynamical conductivities. From (conceptually) simple experiments on these individual conductivities, Hodgkin and Huxley produced simulations that quantitatively described the dynamics of action potentials, showed that the action potentials would propagate along an axon with constant velocity, and correctly described how the velocity should change with axon radius and other parameters. Just as explanations of hydrodynamic phenomena do not require knowledge of the quantum chemistry of water, those who are interested in the behaviour of neural circuits need not know how the particular channel proteins give rise to the Hodgkin–Huxley equations.



Modelling action potentials. The upper trace shows three membrane action potentials, responding to different strengths of stimulus, calculated by Hodgkin and Huxley, while the lower trace shows a corresponding series of experimental recordings. (Adapted from ref. 29.)

Box 2 A decision-making module in bacteriophage lambda



False-colour transmission electron micrograph of lambda bacteriophages ($\times 13,500$).

The bacterial virus lambda can exist in two states inside a bacterial cell. In the lytic state, the virus replicates, producing about 100 progeny virus particles, and releases them by inducing lysis of the host cell. In the lysogenic state, the viral DNA is integrated into the bacterial chromosome and the production of a single viral protein, the repressor, inhibits expression of the other viral genes. The physiology of the host cell and other factors regulate the probability that an infecting lambda virus will become a lysogen, instead of replicating and inducing lysis³⁰.

Elegant phenomenological experiments inferred the existence of bacteriophages and the existence of lytic and lysogenic states³¹ well before the viruses could be seen as physical particles. The isolation of mutants that biased the switch between lysis and

lysogeny defined genes whose products formed part of the switch and sites on the DNA at which these products bound. Sophisticated analysis of the interactions between the mutants led to proposals about the circuitry of the switch, and specific proposals for which DNA sites bound which regulatory proteins. These proposals were verified by molecular analyses that showed that the repressor bound to DNA³² and produced a very detailed description of the biochemical interactions among repressor, other DNA-binding proteins and DNA. Key predictions of models of the switch were verified by reconstructing it in genetically engineered bacteria³³, and by simulating its behaviour using computer models derived from tools used to simulate the behaviour of electrical circuits³⁴.

from other natural sciences and links it to synthetic disciplines such as computer science and engineering.

Is cell biology modular?

A functional module is, by definition, a discrete entity whose function is separable from those of other modules. This separation depends on chemical isolation, which can originate from spatial localization or from chemical specificity. A ribosome, the module that synthesizes proteins, concentrates the reactions involved in making a polypeptide into a single particle, thus spatially isolating its function. A signal transduction system, on the other hand, such as those that govern chemotaxis in bacteria or mating in yeast¹⁻³, is an extended module that achieves its isolation through the specificity of the initial binding of the chemical signal (for example, chemoattractant or pheromone) to receptor proteins, and of the interactions between signalling proteins within the cell. Modules can be insulated from or connected to each other. Insulation allows the cell to carry out many diverse reactions without cross-talk that would harm the cell, whereas connectivity allows one function to influence another. The higher-level properties of cells, such as their ability to integrate information from multiple sources, will be described by the pattern of connections among their functional modules.

The notion of a module is useful only if it involves a small fraction of the cell components in accomplishing a relatively autonomous function. Are modules real? Several lines of evidence suggest that they are. Some modules, such as those for protein synthesis, DNA replication, glycolysis, and even parts of the mitotic spindle (the cellular machinery that ensures the correct distribution of chromosomes at cell division), have been successfully reconstituted *in vitro*. Others are intrinsically more difficult to

reconstruct from purified components and, for these, other methods have established the validity of the module. One method is to transplant the module into a different type of cell. For example, the action potentials characteristic of nerve and muscle cells have been reconstituted by transplanting ion channels and pumps from such cells into non-excitable cells⁴. Another approach is to create theoretical models of the system and verify that their predictions match reality. This approach was used to describe the generation of action potentials long before a molecular description of membrane channels existed (see Box 1). This was the first example of '*in silico* reconstitution', which will have an increasingly important role in cell biology.

Functional modules need not be rigid, fixed structures; a given component may belong to different modules at different times. The function of a module can be quantitatively regulated, or switched between qualitatively different functions, by chemical signals from other modules. Higher-level functions can be built by connecting modules together. For example, the super-module whose function is the accurate distribution of chromosomes to daughter cells at mitosis contains modules that assemble the mitotic spindle, a module that monitors chromosome alignment on the spindle, and a cell-cycle oscillator that regulates transitions between interphase and mitosis.

One must also ask how a cell integrates information and instructions that come from the many different modules that monitor its internal and external environment. Neurobiology has an analogous problem, where the central nervous system integrates information from different senses and dictates the organism's behaviour. Does cellular integration merely emerge from a web of pairwise connections between different sensory modules, or are there specific modules that act as a cellular equivalent of the central

nervous system — integrating information and resolving conflicts?

Complete understanding of a biological module has depended on the ability of phenomenological and molecular analyses to constrain each other (see Box 2). Phenomenological models have fewer variables than molecular descriptions, making them easier to constrain with experimental data, whereas identifying the molecules involved makes it possible to perturb and analyse modules in much greater detail. Thus, the demonstration that genetic information for virulence could be transferred between bacteria prompted the identification of the information-carrying molecule as DNA, before the molecular processes involved in virulence and the structure of DNA were understood. The discovery that genetic information resided in the DNA encouraged structural studies, which then suggested how DNA encodes information and transmits it from generation to generation.

Modular structures may facilitate evolutionary change. Embedding particular functions in discrete modules allows the core function of a module to be robust to change, but allows for changes in the properties and functions of a cell (its phenotype) by altering the connections between different modules. If the function of a protein were to directly affect all properties of the cell, it would be hard to change that protein, because an improvement in one function would probably be offset by impairments in others. But if the function of a protein is restricted to one module, and the connections of that module to other modules are through individual proteins, it will be much easier to modify, make and prune connections to other modules. This idea is supported by the analogous observation that proteins that interact with many other proteins, such as histones, actin and tubulin, have changed very little during evolution,

and by theoretical arguments that proteins are difficult to evolve once they are participating in many different interactions⁵.

Understanding the relatedness of modules is useful because knowledge about one member of a class can inform the study of the others. Relatedness by descent is often apparent from the homology of chemical components. For instance, the mitogen-activated protein kinase cascades that occur in many intracellular signalling pathways define a common functional class of signal transduction modules. Modules may also be related by shared design or functional principles, even if they are not related by descent. The pheromone-detection system of budding yeast and the chemotactic machinery of bacteria use unrelated components, but both pathways achieve a sensitive response over a wide range of pheromone or chemoattractant concentrations by using reactions that specifically turn off active forms of the signalling receptors^{6–8}.

Lessons from other sciences

We have argued that most functional properties of a module are collective properties, arising from the properties of the underlying components and their interactions. Collective properties have long been studied in statistical physics and share attributes that rise above the details⁹. For example, the melting of the surface of a solid can be induced in different ways: by changing the pressure or temperature or by adding impurities. Similarly, different organisms induce the transition between different patterns of microtubule organization that occurs during cell division by changing different members of the set of kinetic parameters that govern microtubule polymerization^{10,11}. The concept of phase (or state) transitions from physics may help unify different observations and experiments. Moreover, many molecular details are simply not needed to describe phenomena on the desired functional level.

Biological systems are very different from the physical or chemical systems analysed by statistical mechanics or hydrodynamics. Statistical mechanics typically deals with systems containing many copies of a few interacting components, whereas cells contain from millions to a few copies of each of thousands of different components, each with very specific interactions. In addition, the components of physical systems are often simple entities, whereas in biology each of the components is often a microscopic device in itself, able to transduce energy and work far from equilibrium¹². As a result, the microscopic description of the biological system is inevitably more lengthy than that of a physical system, and must remain so, unless one moves to a higher level of analysis.

Information flows bidirectionally between different levels of biological organization. For instance, the macroscopic signals

Box 3 From atoms to modules in computers

Building a computer in the 1950s relied on understanding barium oxide, the material of choice for emitting electrons from the cathodes of vacuum tubes. A vacuum-tube module could then be designed whose function was the amplification of a signal, but whose functional description had no reference to barium oxide. These amplifiers were assembled into logical circuits, whose logical operation could be described without reference to vacuum tubes. The connecting wires in these logical circuits had insulation of many colours so that the circuits could be accurately hand-manufactured. Mathematical programs were then designed on the basis of these logic circuits. In the computer of today, there is no barium oxide or coloured wires. Instead, the properties of silicon and silicon dioxide are of primary importance in designing an amplifier, and the transistor replaces the vacuum tube. Both old and new computers have logic circuits based on the same elementary principles, but arranged rather differently as computers have become more sophisticated. Yet all this is unimportant to most users, whose computer program runs on either machine. At one level, barium oxide and coloured wires were the soul of the old machine, while at another level, they are irrelevant to understanding the essence of how a computer functions.



An early stored-program computer (left), built around 1950, used vacuum tubes in logic circuits, whereas modern computers use transistors and silicon wafers (right), but both are based on the same principles.

that a cell receives from its environment can influence which genes it expresses — and thus which proteins it contains at any given time — or even the rate of mutation of its DNA¹³, which could lead to changes in the molecular structures of the proteins. This is in contrast to physical systems where, typically, macroscopic perturbations or higher-level structures do not modify the structure of the molecular components. For example, the existence of vortices in a fluid, although determined by the dynamics of molecules, does not usually change the nature of the constituents and their molecular interactions.

More importantly, what really distinguishes biology from physics are survival and reproduction, and the concomitant notion of function. Therefore, in our opinion, the most effective language to describe functional modules and their interactions will be derived from the synthetic sciences, such as computer science or engineering, in which function appears naturally.

The essence of computational science is the capacity to engineer circuits that transform information from one form into

another on the basis of a set of rules. How might the lessons learned here apply to biology? Evolution selects those members of a genetically diverse population whose descendants proliferate rapidly and survive over many generations. One way of ensuring long-term survival is to use information about the current environment to predict possible future environments and generate responses that maximize the chance of survival and reproduction. This process is a computation, in which the inputs are environmental measurements, the outputs are signals that modulate behaviour, and the rules generate the outputs from the environmental inputs. For example, signals from the environment entrain circadian biological clocks to produce responses to predicted fluctuations in light intensity and temperature. Indeed, the history of life can be described as the evolution of systems that manipulate one set of symbols representing inputs into another set of symbols that represent outputs¹⁴.

Just as electrical engineers design circuits to perform specific functions, modules have

evolved to perform biological functions. The properties of a module's components and molecular connections between them are analogous to the circuit diagram of an electrical device. As biologists we often try to deduce the circuitry of modules by listing their component parts and determining how changing the input of the module affects its output¹⁵. This reverse engineering is extremely difficult. Although an electrical engineer could design many different circuits that would amplify signals, he would find it difficult to deduce the circuit diagram of an unknown amplifier by correlating its outputs with its inputs. It is thus unlikely that we can deduce the circuitry or a higher-level description of a module solely from genome-wide information about gene expression and physical interactions between proteins. Solving this problem is likely to require additional types of information and finding general principles that govern the structure and function of modules.

A number of the design principles of biological systems are familiar to engineers. Positive feedback loops can drive rapid transitions between two different stable states of a system, and negative feedback loops can maintain an output parameter within a narrow range, despite widely fluctuating input. Coincidence detection systems require two or more events to occur simultaneously

in order to activate an output. Amplifiers are built to minimize noise relative to signal, for instance by choosing appropriate time constants for the circuits. Parallel circuits (fail-safe systems) allow an electronic device to survive failures in one of the circuits.

Designs such as these are common in biology. For example, one set of positive feedback loops drives cells rapidly into mitosis, and another makes the exit from mitosis a rapid and irreversible event¹⁶. Negative feedback in bacterial chemotaxis allows the sensory system to detect subtle variations in an input signal whose absolute size can vary by several orders of magnitude¹⁷. Coincidence detection lies at the heart of much of the control of gene transcription in eukaryotes, in which the promoters that regulate gene transcription must commonly be occupied by several different protein transcription factors before a messenger RNA can be produced. Signal transduction systems would be expected to have their characteristic rate constants set so as to reject chance fluctuations, or noise, in the input signal. DNA replication involves a fail-safe system of error correction, with proofreading by the DNA polymerase backed up by a mismatch repair process that removes incorrect bases after the polymerase has moved on. A failure in either process still allows cells to make viable progeny, but simultaneous failure of both is lethal.

In both biological and man-made systems, reducing the frequency of failure often requires an enormous increase in the complexity of circuits. Reducing the frequency at which individual cells give rise to cancer to about 10^{-15} has required human cells to evolve multiple systems for preventing mutations that could generate cancer cells, and for killing cells that have an increased tendency to proliferate.

Biological systems can both resist and exploit random fluctuations, or noise. Thus, evolutionary adaptation depends on DNA being mutable, but because most mutations are neutral or deleterious, the rate of mutation is under rigorous genetic control. Many systems for specifying the polarity of cells or groups of cells rely on a mechanism known as 'lateral inhibition', which causes adjacent cells to follow different fates. This process can amplify a small, often stochastic, initial asymmetry causing adjacent cells or adjacent areas within cells to follow different fates.

Other aspects of functional modules are less familiar to engineers. Several can be subsumed under the idea that the rules for a module's function are rigidly encoded in the structures of its proteins, but produce messy, probabilistic intermediates that are then refined to give unique solutions. This principle seems to hold across an enormous range of scales, from the folding of protein

molecules to the evolution of organisms. The principle arises from a combination of three mechanisms: exploration with selection (trial and error), error-correction mechanisms, and error-detection modules that delay subsequent events until a process has been successfully completed. These are present to different extents in different examples.

Exploration with selection (trial and error) is a fundamental principle of biology and acts on timescales from milliseconds to aeons and at organizational levels from single molecules to populations of organisms. In single molecules, kinetic funnels direct different molecules of the same protein through multiple, different paths from the denatured state to a unique folded structure¹⁸. Within cells, the shape of the mitotic spindle is partly due to selective stabilization of microtubules whose ends are close to a chromosome¹⁹. At the organismal level, the patterning of the nervous system is refined by the death of nerve cells and the decay of synapses that fail to connect to an appropriate target. Within populations, differential reproductive success alters the structure of gene pools, giving rise to evolution.

The use of exploration with selection on short timescales as a design principle in intracellular modules may make them especially easy to modify on evolutionary timescales. The lack of a rigidly programmed

sequence of intermediates should allow such modules to survive incremental modifications and incorporate evolutionary additions such as error detection and correction. Similar messy and probabilistic intermediates appear in engineering systems based on artificial neural networks — mathematical characterizations of information processing that are directly inspired by biology. A neural network can usefully describe complicated deterministic input–output relationships, even though the intermediate calculations through which it proceeds lack any obvious meaning and their choice depends on random noise in a training process²⁰.

Constraints from evolution

One approach to uncovering biological design principles is to ask what constraints they must obey. Apart from the laws of physics and chemistry, most constraints arise from evolution, which has selected particular solutions from a vast range of possible ones.

Today's organisms have an unbroken chain of ancestors stretching back to the origin of life. This constraint has been successfully used to understand protein functions, by comparing existing protein sequences from related species, finding conserved parts and inferring their roles. Comparing modules of common function from different organisms should be a similarly useful tool

for understanding their operation. One has also to remember that today's modules were built by tinkering with already functional modules, rather than by starting from scratch, and may not be the optimal way of solving a particular problem²¹. This evolutionary history is similar to that of man-made devices. Particular solutions in computing, or for any engineered object, are the result of an elaborate historical process of selection by technological, economical and sociological constraints. A familiar example is the less than optimal QWERTY keyboard, originally invented to prevent jammed keys on early manual typewriters. It can be viewed as a living fossil.

The survival of living systems implies that the critical parameters of essential modules, such as the accuracy of chromosome segregation or the periodicity of a circadian clock, are robust: they are insensitive to many environmental and genetic perturbations. Evolvability²², on the other hand, requires that other parameters of modules are sensitive to genetic changes. They can then be modified over many generations to alter the function of a module, or its connections to other modules, in a way that allows organisms to adapt to new challenges. It is important to understand how robustness and flexibility can be reconciled for each functional module.

Organisms have been selected for two properties: rapid reproduction in optimal conditions and the ability to survive rarely encountered extreme conditions. Because environments tend to fluctuate over time, most modules are likely to have been selected for their ability to contribute to both reproduction and survival. These considerations imply that understanding the full function of modules may require us to measure small differences in reproductive ability, as well as studying the performance of modules under extreme perturbations. Some components of an *in vivo* module that are 'nonessential' in normal laboratory conditions are likely to have important roles in the assembly, fidelity, robustness and dynamic characteristics of modules that produce small advantages in long-term survival probability. It may be very difficult, however, to measure such contributions directly.

Survival of a gene pool, as opposed to an individual organism, is favoured by diversification, as the simultaneous presence of multiple phenotypes in a population increases the possibility that some individuals will survive and reproduce in a heterogeneous and changing environment. Diversification can be achieved by epigenetic mechanisms that enable a single genotype to produce more than one phenotype, by genetic mechanisms that maintain multiple genotypes in a population, and by speciation, which splits a single gene pool into two independently evolving pools. It is thus important to consider the function of modules not only in the context of an organism, but also from the point of view of the population of organisms, and to ask how modular construction facilitates the maintenance and selection of diversity.

Towards modular biology

A major challenge for science in the twenty-first century is to develop an integrated understanding of how cells and organisms survive and reproduce. Cell biology is in transition from a science that was preoccupied with assigning functions to individual proteins or genes, to one that is now trying to cope with the complex sets of molecules that interact to form functional modules^{23,24}. There are several questions that we want to answer. What are the parts of modules, how does their interaction produce a given function, and which of their properties are robust and which are evolvable? How are modules constructed during evolution and how can their functions change under selective pressure? How do connections between modules change during evolution to alter the behaviour of cells and organisms?

The number of modules that have been analysed in detail is very small, and each of these efforts has required intensive study. Biologists need to study more functions at the modular level and develop methods that make it easier to determine the relationship

of inputs to outputs of modules, their biochemical connectivity, and the states of key intermediates within them. Three complementary approaches can help in this task: better methods for perturbing and monitoring dynamic processes in cells and organisms; reconstituting functional modules from their constituent parts, or designing and building new ones; and new frameworks for quantitative description and modelling of modules.

The first approach is illustrated by efforts to find small organic molecules that can perturb and report on the activity of modules. Calcium-binding dyes have been used to follow the activities of neurons with high spatial and temporal resolution. Light-activated chemicals can perturb function on the timescales that characterize changes within the modules, thus giving them an important advantage over the slower perturbations produced by classical and molecular genetics²⁵. Another example of a new method for monitoring cellular processes is genome-wide analysis of gene expression^{26,27}. But techniques for collecting information about the entire genome will be only as powerful as the tools available to analyse it, just as our ability to infer protein structure and function from protein sequence data has increased with the sophistication of tools for sequence analysis. We need better methods of finding patterns that identify networks and their components, of identifying possible connections among the components, and of reconstructing the evolution of modules by comparing information from many different organisms.

Another approach to discerning module function is that of 'synthetic biology'. Just as chemists have tested their understanding of synthetic pathways by making molecules, biologists can test their ideas about modules by attempting to reconstitute or build functional modules. This approach has already been used to construct and analyse artificial chromosomes made by assembling defined DNA elements²⁸, and cellular oscillators made from networks of transcriptional regulatory proteins (M. Elowitz and S. L., unpublished results). Seeing how well the behaviour of such modules matches our expectations is a critical test of how well we understand biological design principles.

The main difficulty in reconstructing the evolution of modules is our ignorance about past events. One solution to this problem is to examine the evolution of module function in the laboratory. Analysing multiple repetitions of such experiments may tell us how much the path of future change is restricted by the current structure and function of modules, and should help us to understand how evolutionary pressures constrain biological design principles.

Finally, we emphasize the importance of integrating experimental approaches with

modelling and conceptual frameworks. The best test of our understanding of cells will be to make quantitative predictions about their behaviour and test them. This will require detailed simulations of the biochemical processes taking place within the modules. But making predictions is not synonymous with understanding. We need to develop simplifying, higher-level models and find general principles that will allow us to grasp and manipulate the functions of biological modules. The next generation of students should learn how to look for amplifiers and logic circuits, as well as to describe and look for molecules and genes (Box 3). Connecting different levels of analysis — from molecules, through modules, to organisms — is essential for an understanding of biology that will satisfy human curiosity.

Leland H. Hartwell is at the Fred Hutchinson Cancer Center, Seattle, Washington 98109, USA. John J. Hopfield and Stanislas Leibler are in the Department of Molecular Biology (J.H.) and Department of Physics and Molecular Biology (S.L.), Princeton University, Princeton, New Jersey 08542, USA. Andrew W. Murray is in the Department of Physiology, University of California at San Francisco, San Francisco, California 94143, USA.

1. Stock, J. B. & Surette, M. G. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds Neidhardt, F. C. & Curtiss, R.) 1103–1129 (ASM Press, Washington, DC, 1996).
2. Herskowitz, I. *Cell* **80**, 187–197 (1995).
3. Posas, F., Takekawa, M. & Saito, H. *Curr. Opin. Microbiol.* **1**, 175–182 (1998).
4. Hsu, H. *et al. Biophys. J.* **65**, 1196–1206 (1993).
5. Waxman, D. & Peck, J. R. *Science* **279**, 1210–1213 (1998).
6. Konopka, J. B., Jenness, D. D. & Hartwell, L. H. *Cell* **54**, 609–620 (1988).
7. Goy, M. F., Springer, M. S. & Adler, J. *Proc. Natl Acad. Sci. USA* **74**, 4964–4968 (1977).
8. Barkai, N. & Leibler, S. *Nature* **387**, 913–917 (1997).
9. Anderson, P. W. *Science* **177**, 393–396 (1972).
10. Belmont, L. D., Hyman, A. A., Sawin, K. E. & Mitchison, T. J. *Cell* **62**, 579–589 (1990).
11. Gliksmann, N. R., Parsons, S. F. & Salmon, E. D. *J. Cell Biol.* **119**, 1271–1276 (1992).
12. Alberts, B. & Miake-Lye, R. *Cell* **68**, 415–420 (1992).
13. Shapiro, J. A. *Ann. NY Acad. Sci.* **870**, 23–35 (1999).
14. Hopfield, J. J. *J. Theor. Biol.* **171**, 53–60 (1994).
15. Bray, D. *Nature* **376**, 307–312 (1995).
16. Morgan, D. O. *Annu. Rev. Cell. Dev. Biol.* **13**, 261–291 (1997).
17. Berg, H. C. *Cold Spring Harb. Symp. Quant. Biol.* **53**, 1–9 (1988).
18. Dill, K. A. & Chan, H. S. *Nature Struct. Biol.* **4**, 10–19 (1997).
19. Heald, R. *et al. Nature* **382**, 420–425 (1996).
20. Sejnowski, T. J. & Rosenberg, C. R. *Complex Systems* **1**, 145–168 (1987).
21. Jacob, F. *Science* **196**, 1161–1166 (1977).
22. Kirschner, M. & Gerhart, J. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427 (1998).
23. Nurse, P. in *Limits Of Reductionism In Biology* Ciba Foundation Symp. **213**, 93–101 (1998).
24. Strohmman, R. C. *Nature Biotech.* **15**, 194–200 (1997).
25. Adams, S. R. & Tsien, R. Y. *Annu. Rev. Physiol.* **55**, 755–784 (1993).
26. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. *Science* **270**, 467–470 (1995).
27. Brown, P. O. & Botstein, D. *Nature Genet.* **21**, 33–37 (1999).
28. Murray, A. W. & Szostak, J. W. *Nature* **305**, 189–193 (1983).
29. Hodgkin, A. L. & Huxley, A. F. *J. Physiol.* **117**, 500–544 (1952).
30. Herskowitz, I. & Hagen, D. *Annu. Rev. Genet.* **14**, 399–445 (1980).
31. Lwoff, A. *Bact. Rev.* **17**, 269 (1953).
32. Ptashne, M. *Nature* **214**, 232–234 (1967).
33. Ptashne, M. *et al. Cell* **19**, 1–11 (1980).
34. McAdams, H. H. & Shapiro, L. *Science* **269**, 650–656 (1995).

Flashes in femtoseconds

For the molecule, time ticks in picoseconds, or even femtoseconds — respectively 10^{-12} and 10^{-15} s. On such timescales, a molecule tumbles through space, an atom vibrates back and forth in a crystal, a chemical bond is forged or broken. Yet the development of pulsed lasers blinking in a femtosecond staccato of flashes has enabled these fundamental molecular processes to be charted, revealing the rise and decline of ephemeral intermediates under the furiously strobed illumination.

Femtosecond spectroscopy has shown us the dynamics of bond breaking — but what of the structural aspects? When will we be able to map out the coordinates of individual atoms, as, for example, a protein docks onto the cell surface or as a chlorine radical eliminates an ozone molecule? When will we see the first atomically resolved diffraction pattern of a true transition state, and watch frame by frame as vibrations carry atoms into new unions?

This demands a marriage of the techniques of ultrafast spectroscopy with those of X-ray diffraction. The probe beam becomes a pulsed X-ray source, its flashes brief enough to 'freeze' the atomic motions yet bright enough to provide a discernible diffraction pattern. Current synchrotron sources can be pulsed at around 50 picoseconds, whereas ultrashort laser pulses can excite bursts of X-rays from suitable targets that could, in principle, be over within 100 femtoseconds.

These pulses must be synchronized with an optical pumping beam, which gets the reaction underway. An alternative is to diffract electrons rather than X-rays. Diffraction on the 'molecular' timescale of femtoseconds is an infant discipline which promises wonders once perfected, but which is capable right now of only the crudest of impressionistic sketches: blurred images of lattice dynamics, showing evidence of rapid change but without a single molecule (let alone an atom) in focus. The static photography of the Braggs has yet to produce its first movie. **Philip Ball**

Arrhythmias

'His heart raced'. 'Her heart skipped a beat'. 'My heart beat sluggishly in my breast.' Literature is peppered with such lines. Why? Because, as many a dramatic writer has noted, a disturbance in that most basic 'rhythm of life' — the normal human heart rate of between 60 and 100 beats per minute — is a sign that something is amiss.

Normally a heartbeat starts in the right atrium of the heart, in a specialized group of pacemaker cells known as the sinus node. This node sends an electrical signal, which spreads through the atria and ventricles of the heart, following a precise route that induces the heart to beat in the most efficient way.

If the sinus node develops an abnormal rate or rhythm, if the conduction pathway of its electrical signals is faulty or if another part of the heart takes over as the pacemaker, arrhythmias — as the various types of irregular heartbeat are generically known — can ensue. Arrhythmias may be due to heart disease, ageing, medications, metabolic imbalances or other genetic or infectious medical problems. And the ineffective blood pumping that they give rise to can cause fatigue, dizziness, light-headedness, fainting spells and, ultimately, strokes and death.

Oscillations in the cells of the sinus node operate by setting up a cyclic decay of the difference between the charge on the outside of the membrane and the inside. This causes the membrane to fire spontaneously, which raises its conductance and allows electrolytes to rush in, thus reversing the potential across the membrane. This reversed charge propagates to neighbouring cells, causing the same sequence of events therein and hence a concert of muscular contraction.

But it is currently unclear exactly how cyclically expressed genes and proteins give rise to cyclic decay in membrane potential in the heart's muscle cells. So, just as in many other areas of medical biology, the challenge now is twofold. First, to expand our knowledge of the genetics and cell signalling processes that gives rise to healthy heartbeats. And second, to integrate this into a far richer picture of the workings, and failings, of the whole heart.

Sara Abdulla

Does the past have a future?

Hardly a week goes by without the discovery of some amazing fossil that will change the way we look at the history of life. **But how long can the fossil bonanza last?** After all, the Earth has only so many rocks to investigate. Will there come a time when we can say that the fossil record is substantially complete, in that the likelihood of discovering radically new forms, or significant range extensions of known forms, becomes improbable? If so, when will that time arrive? And when it does, how will this influence the practice of palaeontologists?

"I think we are more or less at that point now," says Andrew Smith of the Natural History Museum in London. "The available outcrops are pretty heavily searched within the developed nations, and major headway has been made in exploring places like Mongolia and Madagascar." His prognosis is, however, tempered by the fact that absolute completeness is an impossibility, because the vast majority of organisms do not end up as fossils: **"my belief is that we will have a substantially biased fossil record that cannot be bettered no matter where we go in the world".**

Charles Marshall of Harvard University agrees that the Earth's fossil riches have been fairly well worked out, but says that one cannot discern a point when the last fossil will have been extracted. "One can't simply extrapolate from what has been discovered because radical new finds often result from unanticipated ways of looking, often in stratigraphic intervals or rock types not typically examined. If the rate of recent discoveries, from fossil embryos to down-covered dinosaurs, are any indication there is much to be discovered yet!"

However, contrary to popular iconography, most palaeontologists do not devote most of their time to making radical discoveries. Such discoveries, says Marshall, will be — as they are now — "largely the unexpected by-products of mature research programmes designed to establish the patterns of evolution and elucidate the processes responsible for those patterns."

"The obvious priority," says Smith, "will be to establish and understand how sequence stratigraphy and sea-level change control facies preservation, and thus control apparent biotic diversity and extinction patterns through time. Only then will we begin to get a reliable handle on how diversity has changed over time."

Henry Gee

Keeping time

How finely can time be measured? The answer might define our capacity to conduct precise tests of general relativity, as demonstrated already in the observed increase in clock rate with distance from the Earth's surface, or the measurement of relativistic effects in binary pulsar systems. Spacecraft navigation, interferometric radio astronomy and Global Positioning System geodesy are all dependent on timekeeping that taxes current technology. And accurate measurements of the second define other standards, such as the units of distance and voltage.

Measuring time is about monitoring oscillations, whether of the pendulum, the quartz crystal or the caesium atom. By relying on atoms for our frequency (and hence time) standards, we may at least avoid placing ourselves at the mercy of the machinist who made the device; but on the other hand, we fall foul of the Uncertainty Principle. Atomic transitions, of which the hyperfine splitting of energy levels in caesium-133 furnishes us with today's standard second, happen at a frequency that can ultimately be well defined only to the degree that the excited state is long-lived. The shorter the lifetime — or the duration of the measurement — the greater the uncertainty in energy (and frequency) of the transition.

So to get a more accurate result, one needs a more leisurely attitude to measurement. This is the principle behind the 'atom fountain', a device that first cools atoms within an optical 'molasses' of three crossed laser beams before gently launching them vertically so that they fall back under gravity. The atoms can then be probed outside the perturbing influence of the laser field.

Towards the top of the fountain, the atoms pass twice through a microwave cavity: once on the way up, once on the way down, separated by about a second. Invented by Norman Ramsey in 1949, this 'separated oscillatory-field method' is applied to a linear atomic beam in current atomic clocks. But the atom fountain permits a much longer span between pulses, and so a greater narrowing of the transition. In principle these devices offer a frequency determination accurate to one part in 10^{16} , which is equivalent to a variation of about one second in a billion years.

Philip Ball

Circadian clocks

How do organisms respond to their most fundamental timescale — the Earth's rotation around its own axis? With a suite of interconnected hormonal and physiological loops that make up the 'circadian timing system' or CTS.

Every CTS — whatever the organism — is made up of light receptors, an endogenous 'biological' clock and a so-called 'entrainment' system to couple one with the other. For although every cell of every biological or circadian clock behaves cyclically, these intrinsic periods are not necessarily exactly 24 hours long. The clock has to be set.

Much of the pathology and endocrinology of the circadian clock has been elucidated in the past century. In humans, for instance, many sections of the route from retina to brain to heart-muscle action, metabolism, body temperature and the organization of sleep-wake cycles have been charted. But there are still major areas of uncertainty, **especially in the newest branch of circadian biology: the identification and study of genes necessary for normal timekeeping.**

Clock genes seem to be present in all cells, even though only a subset is capable of acting as self-sustained oscillators. So do these genes have other, global, non-circadian functions in cell biology? As Russell Foster of Imperial College, London, puts it: "cellular assays have shown the way forward, now we need good functional analysis using transgenic approaches and behavioural studies".

Moreover, only a few animal systems have been studied so far in any detail — the mouse, the fly and the fungus, *Neurospora*. This has revealed some remarkable similarities but also some exciting differences. Comparisons of additional species will surely tell us more about the evolution and adaptive significance of circadian systems and their interactions with their environments. All of which will hopefully feed back into a fine-tuned understanding of, and ability to treat or even to manipulate, the one timepiece that we all care about — the human body.

Sara Abdulla

The challenge of conservation

Some disciplines move faster than others, but if there is one moving much faster than its practitioners would like, it is conservation biology. In the next century, conservationists will be less interested in pristine wilderness — for such will hardly exist — but in the success with which human beings manage the biosphere to the benefit of all its inhabitants, including humans. The world is currently in an unprecedented situation in which *Homo sapiens* — just one species out of millions — sequesters 45 per cent of the planet's net terrestrial productivity and 55 per cent of its fresh water. Effective biodiversity systems management will become an imperative if only to maintain the quality of human life, let alone that of other species.

In the coming decades, conservation biology will become a close analysis of the art of compromise, as humanity seeks to understand the global ecosystem as a participant, not as an observer. "Within a century," says Peter Raven, director of the Missouri Botanic Garden, "the world will be a patchwork with varying degrees of biological richness, and the way people respond locally to challenges will determine the nature of that compromise."

What will we have lost by 2100? Raven estimates that two-thirds of all species will disappear if current trends continue, but is nevertheless optimistic. "We can do better," he says. **"No major category of organisms, or local community of organisms, need disappear completely."**

Ecologists are acutely aware that the maintenance of ecosystem function transcends the conservation of any particular species, so high-technology initiatives such as 'ex situ' conservation — maintaining banks of germplasm outside the context of a habitat — may become a side issue.

"High technology is unlikely to solve most of the major global environmental issues that we face,"

says David Tilman of the University of Minnesota. "Rather, lifestyle changes, most likely imposed by legislation, will be needed. Such legislation will be socially sustainable only once citizens understand their long-term dependence on biodiversity and ecosystem services." And Raven adds that it is "better for now to follow Aldo Leopold's dictum: 'the first rule of intelligent tinkering is to save all the cogs and wheels'".

Henry Gee



Physics at the Planck time

Superstring theory, perhaps the most notorious attempt to unify general relativity with quantum mechanics, was once described as a piece of twenty-first-century physics that fell into the twentieth century. Certainly, physicists struggling to marry these reluctant partners have long sensed that the crucial piece of the puzzle has yet to be invented. But a theory of quantum gravity remains more a question of tidiness than anything else — quantum effects on space-time become important only on time and length scales well out of experimental reach. A quantum field theory of electromagnetism has plenty of practical relevance for applied physics, and such a field theory for the nuclear forces generates no shortage of testable predictions. But gravity yields to the quantum rod only at the so-called Planck scale: over distances of about 10^{-33} cm and times of 10^{-43} s.

A physics to match the Planck timescale is the biggest challenge to physicists in the coming century. **This is the timescale relevant to the graviton, the putative carrier of the gravitational force.** And naturally enough, only such a theory can extend our understanding of the Big Bang inside the first 10^{-43} seconds of existence.

What will a theory of quantum gravity do to space-time? Will it be like John Wheeler's 'quantum foam', furiously alive with ephemeral black holes and worm holes? Or will it be more like Abhay Ashtekar's fabric of woven loops? Will particles become strings, and will the strings be supersymmetric? How many extra dimensions will turn up, curled out of sight over the Planck distance?

And will we ever be able to test such a theory? One quantum-gravity researcher has confessed that "it seems highly unlikely that a machine will ever be built with which these minute distances can be studied directly". For guidance, it seems that physicists may have to fall back on the old stand-bys: elegance and beauty. **Philip Ball**

The speed of computers

One version of Moore's Law, predicted by Intel co-founder Gordon Moore in 1965, has it that computer chips double in speed every 18 months. Nothing moves any faster, of course — the chips simply become more powerful by packing more devices into the same space. Speed is about miniaturization. **Recent research on the ultimate size limits for silicon technology have fuelled speculation about whether Moore's Law is headed for a crash.** By 2012, at the current rate of shrinkage, transistors will have become so small that the silicon oxide films will no longer guarantee adequate insulation between conducting regions: they become switches that cannot be fully turned off. What happens then to our expectation that the next generation of machines will be better, meaner and faster?

All manner of factors determine how quickly, say, the *Nature* web page downloads onto your screen, many of them scarcely connected to the component density on the chips inside the processor. In some parts of the world it can take ten minutes or more. Yet you would be waiting a lot longer if the data were streaming down old copper telephone cables rather than glass optical fibres. Fibre-optic networks have a greater carrying capacity that speeds up transmission by one or two orders of magnitude relative to electrical lines. But that is only a fraction of what should be possible with photonic technology.

And the electronics at either end are soon to be stretched to their limit to cope with the transmission speeds that photonics offers. Gigabit-per-second distributed networks have already been demonstrated, and laser diodes can in principle blink out 100 Gbits per second; but electronics lose the ability to cope at half that rate. **All-optical networks will then be required, and networks working at 100 Gbits per second are already in development.** At some point, the light will need to go on a chip, in photonic integrated circuits.

But photonics is an embryonic technology, and plagued with obstacles in comparison with mature electronics. No clear rival to the transistor has yet emerged from this or any other direction, such as single-electron devices and molecular electronics. So perhaps we will have to adjust our expectations as the information revolution reaches a plateau. Moore predicts that we have about two generations of computers left with current technologies, and "beyond that, life gets very interesting". **Philip Ball**

Lifespan extension

Eat more fibre! Consume less cholesterol! Drizzle more olive oil! Have fewer children later in life! Drink a glass of red wine a day! And so it goes on. Rarely a day goes by without the emergence of a new 'theory' in that most seductive realm of scientific advances — lifespan extension.

Chromosome structure hints that our cells have genetic 'fuses' that mete out their allotted time. Antibiotics, vaccinations and antiseptics have brought about hikes in life expectancy in the developed world which have raised the average lifespan to almost double that at the last *fin de siècle* — nigh-on 80 years and still rising — as this century draws to a close. The intractability of several cancers such as Hodgkin's lymphoma and leukaemia has been tamed and many others look to follow soon; the hard nut of cognitive decline is on the verge of being cracked with drugs and neuron replacement; and workable synthetic organs and xenotransplants are around the corner.

But the current obsession with all things genetic, molecular, high-tech and expensive risks missing the point that both the World Bank and World Health Organization (WHO) make patently clear. That is, **for the majority of the planet the most powerful lifespan-extending technologies of all — good, basic public health and preventative medicine — are still a far-off dream, even if they are, scientifically speaking, old hat.**

As we usher in the third millennium, the WHO estimates that the citizens of the world's poorest countries (who make up two-fifths of the global population) have almost half the life expectancy — at under 50 years — of those living in the richest nations. And that gap is widening, not narrowing,

often in both directions. One third of the planet's population do not have secure access to safe water or sufficient food, let alone drugs, immunizations, or even the simplest health education. Elsewhere, billions of dollars are being spent teasing out the intricacies of gene therapy and memory enhancement. Addressing this inequity is surely one of the greatest and most pressing challenges facing us all, if not as academics, then as people. **Sara Abdulla**

Computing 2010: from black holes to biology

Declan Butler

By 2010, a click on the PC on your desktop will suffice to call up instantly all the computing power you need from what by then will be the world's largest supercomputer, the Internet itself. Supercomputing for the masses will trigger a revolution in the complexity of problems that are tackled, whole disciplines will go digital and, rather than spending time collecting their own data, scientists will organize themselves around shared data sets.

"PAL, I'm ready; set up that conference call." The teraflop PC in George's Chicago laboratory obeys. Detectors stir and projectors purr. Ceiling, walls and floor shimmer sapphire. The virtual-reality cavern ready, PAL connects to the Internet at ten gigabytes per second, beaming in three-dimensional avatars of Juliette in Paris and Günther in Munich. The three shake hands. The experiment begins.

Elsewhere on the Internet, PAL is busy interpreting and assembling data sets from research centres across the globe, while simultaneously negotiating CPU time from a supercomputer in New Mexico, and petaflop clusters of PCs in London and Prague. Operation accomplished. A kaleidoscope cloud of millions of glowing pixels suddenly coalesces in the centre of the laboratory, sculpting two floating black holes that swirl slowly around one another.

Koichi, the principal investigator, takes control of the input parameters, steering a collision of the black holes using a handheld screen from his bedroom in Tokyo. One of the most violent events in the Universe unfolds in fast-forward in front of the scientists, as the black holes whirlpool in the final plunge, merge and mushroom.

Within minutes, PAL has crunched the terabytes of output. "Back in the 1990s, it would have taken months," muses George to himself. The results scroll down the high-resolution

wall. An unequivocal first; a strong burst of gravitational waves in the last milliseconds of the collision, exactly as predicted by Einstein's theory of relativity.

The excitement builds. PAL relays even better news — that the bursts also have a specific waveform — to its counterpart machine at VIRGO, the European gravitational-wave detector in Cascina, Italy. VIRGO had come on-line in 2008, and its kilometres-long laser interferometers had amassed a haystack of data; but finding the needle — tell-tale vibrations a tiny fraction of an atomic nucleus in magnitude — had proved elusive. VIRGO now had a vital clue as to what to look for.

The scientists pace for what seems an eternity, to be interrupted by on-line gatecrashers from VIRGO, popping champagne. PAL's waveform template matches blips buried in the noise of VIRGO's massive data sets. The signal-to-noise ratio adjusted, the blips even yield a first guess at the mass of a pair of real black holes. PAL generates a virtual draft manuscript, which is passed around, quickly edited and passed back. The computer rapidly ranks the citation and media impact of the publication in the EMBIH global literature repository as well as the handful of journals that now exist, and — showing signs of intelligence — decides to e-mail the paper to Nature. All in a day's work in 2010.

Journalistic licence? Certainly. Fiction? Probably not. The core science of the black-hole collision, and much of the Internet technology, was demonstrated in June this year, when a team led by Ed Seidel of the Albert Einstein Institute in Potsdam, and Wai-Mo Suen of Washington University in St Louis, broke supercomputing records with a 140,000-CPU-hour run that produced almost a terabyte of data, using the 256-processor 'Origin2000' at the US National Center for Supercomputing Applications (NCSA).

Prototype systems emerging from the world's most advanced computing and networking test beds suggest that such high-end

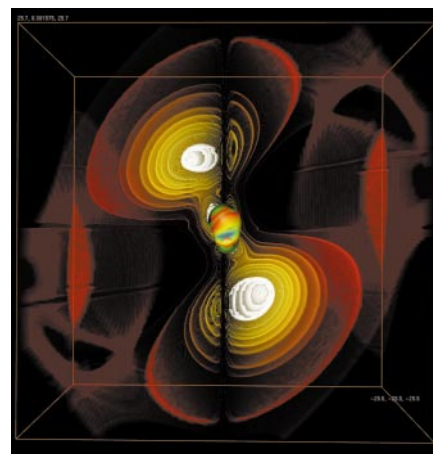
virtual reality, modelling and visualization will be routine over tomorrow's Internet. A proprietary virtual-reality theatre, CAVE, under development at the Electronic Visualization Laboratory of the University of Illinois at Chicago, is one of several systems for remote collaboration — or 'tele-immersion' — that are up and running. Nor will using such technologies require booking scarce run-time on supercomputers in advance; a click on the PC on your desktop will suffice to call up instantly all the computing power you need, from what by then will be the world's largest supercomputer, the Internet itself.

By 2010, today's top-end computing promises to be no longer the exotic pastime

of a privileged few, but the routine of the masses. A vision is emerging of a global computing grid, where computers of any shape and size are linked across the Internet to assemble giant distributed supercomputers. Scientists in every discipline will, from their desktops, have access to raw computing power that dwarfs today's most powerful supercomputers.

Seidel admits that simulations of black-hole collisions are still far from being sophisticated enough to provide realistic help to the planned European VIRGO (Variability of Solar Irradiance and Gravity Oscillations) and US LIGO (Laser Interferometer Gravity-wave Observatory) gravitational-wave observatories. But he predicts that they soon will be, as advances in computing power and algorithms enable such complex simulations to model reality for the first time.

The wide availability of massive comput-



Black-hole simulation showing a just-formed larger black hole that resulted from the collision of two smaller black holes. The surface of the newly formed hole is shown as a coloured surface in the centre of the simulation. This surface is semi-transparent, allowing one to see the two original holes inside. A burst of gravitational waves results from this collision, and these waves are shown as the red-yellow wisps emanating from the central region.

AEI/NCSA/UNIV. WASHINGTON

ing power to help tackle complex systems will prompt a profound change in science itself, predicts Ruzena Bajcsy, assistant director of the US National Science Foundation (NSF)'s directorate for computer and information science and engineering. She believes it will close the chapter of Cartesian reductionism in research, and usher in a new era where researchers will increasingly take on the understanding of complex dynamic systems, such as whole cells and the Earth. To advance the global grid, NSF is funding prototypes for scientific use — through the National Partnership for Advanced Computational Infrastructure, a consortium led by the San Diego Supercomputer Center — and for the population at large — through the National Computational Science Alliance, an NCSA-led consortium of over 50 research agencies and universities.

Conventional supercomputers will without doubt become major nodes on the grid at least at the beginning, and linking even a few of these would greatly increase the resources that could be brought to bear on individual problems. Petaflop supercomputers that are roughly 1,000 times faster than today's top teraflop (10^{12} floating-point operations per second) machines are expected soon after 2010 — and perhaps before if current research on technologies such as 'processor in memory' and the use of superconducting parts bears fruit.

But supercomputers are expensive. The world's most powerful supercomputer, Asci Red, at Sandia National Laboratories in New Mexico, which distributes tasks across an array of around 10,000 Pentium Pro chips, cost almost US\$60 million to build. A decade ago, Tom Sterling, who works at Caltech and NASA's Jet Propulsion Laboratory, thought that a cheaper way forward might be to simply string together lots of PCs. Sterling built his first 'Beowulf' cluster supercomputer by taking a then-barely-heard-of open-source operating system, Linux, and adding network drivers to it. The project worked, cost a tenth of the price of a conventional supercomputer and was also highly flexible — if more power was needed then more PCs were added. 'Avalon', a Beowulf at the Los Alamos National Laboratory, ties together 140 PCs and carries out 50 billion floating-point calculations per second for a construction cost of a mere \$300,000. The European Southern Observatory at La Silla, Chile, and LIGO have also chosen Beowulfs to meet their supercomputing needs in data processing

The 64-bit question

The Beowulf concept now looks set to take off in an even bigger way: on the Internet. Intel's latest chip, the IA64, or Merced — which is due to be shipped in 2000 — is more than just another faster Pentium. What makes Merced so special is that it is the first commodity PC chip to have a 64-bit architecture.



CAVE virtual-reality theatre.

Many scientific applications and high-resolution graphics require 64-bit floating-point accuracy — today's 32-bit PCs round numbers off too imprecisely for many tasks. The arrival of the 64-bit chip blurs further the differences between Beowulfs and conventional supercomputers. Being able to link up cheap 64-bit PCs for scientific purposes will open up cluster supercomputing to many more laboratories, while the movement will also be encouraged by the arrival of gigabit-per-second Ethernet local area networks.

The designers of tomorrow's Internet think they might be able to take the idea further. If you can combine the computing power of PCs over a local network, then why not also over the high-speed network that will be tomorrow's Internet? 'Asci Red has 10,000 Pentium Pros; link up PCs on the Internet and you could link up tens of millions,' says Larry Smarr, director of the National Computational Science Alliance. 'We will have the accuracy we need for these large complex simulations but at the price of PCs'.

Indeed, the characteristics of 64-bit chips, in principle, make them good candidates for such a feat. Although today's 32-bit PC chips can address 2^{32} other items, which in terms of memory represents 2 gigabytes, 64-bit chips can address 2^{64} , which in distributed computing terms means that one processor could in principle address the memory of every other computer on the planet. Link such PCs over a high-speed Internet and it would in effect become one

large computer, says George Lake, a scientist at the University of Washington and NASA project scientist for high-performance computing in Earth and space science. (The 128-bit chips on the horizon would allow the user to address a number equivalent to every atom in the Universe with space left over.)

'Think ahead ten years,' says Smarr, 'and you can see the supercomputer sort of dissolving into this vast planetary fabric of computing because all [computer technology] will be built on the same commodity processors, operating systems and interconnects.'

Proof of principle for the use of PCs over the Internet comes from the Search for ExtraTerrestrial Intelligence screensaver SETI@home, a downloadable screensaver which has become the largest distributed computing project in history. Over 1 million users (including this author) have donated spare CPU time on their PCs and Macs to search for signals from extraterrestrials in the 35 gigabytes of data produced daily from the Arecibo radio telescope in Puerto Rico. The result is the planet's largest virtual supercomputer, running at 7 teraflops.

David Anderson, a computer scientist at the University of California, Berkeley, who leads the project, says that he now intends to use the same approach to tackle other supercomputing tasks including rational drug design, protein folding and high-resolution graphics. A web-based group, 'distributed.net', has also made a name for itself by using distributed computing over PCs to solve cryptography keys. But Sterling

remains cautious about the prospect of commodity PC chips becoming a major source of supercomputing power over the Internet, despite its extraordinary potential. He points out that the speed at which the memory is addressed — or latency — is crucial for many applications. At present, the fast interconnects needed are found only in dedicated supercomputers. The connections between clusters of PC operating over local area networks are still too slow or unreliable for applications requiring tight coupling, let alone clusters operating over the network.

Faster interconnects produced by Gigaset Inc. have produced tighter coupling in a 256-processor cluster unveiled in October by the Advanced Cluster Computing Consortium, an organization based at Cornell University which includes Microsoft, Dell and Intel. The project, based not on Linux but on Windows 2000, is a commercial bid to standardize reliable cluster computing systems for supercomputing in research and business.

Latency is a wider problem in computing. The time required to fetch something from the memory of say a Cray supercomputer — 300 nanoseconds — is not much less than half what it was 20 years ago. Ultimately, this means that significantly greater speeds will not necessarily follow from adding more transistors on a chip. Putting the central processing unit in the memory is the only feasible option, according to many computer scientists.

What seems clear is that as soon as the high-speed Internet becomes available, distributed computing in some form is going to come of age, probably involving a mix of supercomputers, clusters of PCs and workstations, and individual PCs. The backbone of the US research network currently runs at 2.5 gigabytes per second. But Douglas E. Van Houweling, president of the University Corporation for Advanced Internet Development (UCAID), which runs the US Internet 2 project, predicts that the network will run at terabyte levels by 2010, delivering to gigabit desktop connections.

Power hungry

Brute-force computing is badly needed. "With 100,000 sequences of 600 bases pouring out every day and needing to be screened against all others, we could easily use ten times the computing power right now just for sequencing," says Craig Venter, chief executive officer of Celera Genomics Corporation. "If you do not have this kind of [high-end] computing capacity in biology, tomorrow you are going to get left behind."

Demand for supercomputing power at the NSF is growing exponentially and will pass 10 teraflops by 2005, with most coming from new users. The National Computational Science Alliance has 1,536 processors for scientific supercomputing, but Smarr points out that with around 300 users a month, that

works out at a handful of processors per user. "The peer-review system to select projects is a polite term for 'rationing,'" quips Smarr. Over half the top projects get rejected but Smarr is optimistic that over the next decade we will move from "an era of scarcity to an era of plenty."

Attention is now turning to how best to make such distributed resources available to users. The possibilities are enormous. Imagine sitting at your screen in front of a simple Yahoo-type interface, customized to your personal research requirements. One click and databases around the world are scoured for the data you want, say spectral lines of several classes of stars. Little matter what format the image bank databases are in; the interface uses a standard format and automatically converts data having other formats. With the data collected and merged, another click and the interface negotiates with computers around the world to book in real time all the CPU time you need. Want a Fast Fourier Transform? One keystroke does it. If you wish to view the results in three dimensions, virtual reality or rotate them, click on a menu offering advanced visualization techniques.



Larry Smarr foresees supercomputers being superseded by a vast planetary fabric of computing.

And all translated into your own native language.

Such interfaces to a suite of tools written specifically to exploit the distributed database of the Internet and invisible to the user — or 'middleware' — are now the focus of development within many advanced Internet engineering projects such as Internet 2. By 2010 they will be a routine part of the infrastructure of working scientists, according to Rick Stevens, head of the mathematics and computer science division at Argonne National Laboratory.

Plug and play supercomputing

A big step in this direction is the recent release of Cactus, a novel software tool kit developed by Joan Massó and Paul Walker, at the Max Planck Institute for Gravitational Physics in Potsdam. It allows scientists to combine supercomputers over a high-speed network without having to know anything about the necessary advanced computing techniques. It is plug and play in that users simply take their own application (or 'thorn') — be it written in Fortran or C or whatever — and plug it into the flesh of the Cactus code. The Cactus automatically 'parallelizes' their program to run on virtually any computer system, from a portable PC to

distributed clusters of PCs or supercomputers. Cactus's modules also allow booking of CPU time at remote centres, handle the output of terabytes of data, and also offer a suite of state-of-the-art three-dimensional visualization tools. The NSF has committed \$2.2 million to developing Cactus in the United States. The code is accessible to the public at <http://www.cactuscode.org>.

That is a sea change. In the past, supercomputing algorithms have generally been developed for specific problems, and therefore could not be used by others. Cactus will allow biologists, for example, to draw on decades of experience of their more mathematically inclined colleagues, physicists. "As the network and network software evolves I think the division between the PC on your desk and the assets the networks will offer will become steadily less apparent," predicts Van Houweling. "Today, scientists mainly use the web to find information. In the future, accessing large-scale computing resources over the net will be the major use."

Such 'middleware' is beginning to appear in biology. Even the most basic genomic analyses often require users to run a plethora of different software and data formats. NCSA's 'Biology Workbench' and Baylor College of Medicine's 'Search Launcher' are two prototype systems designed to replace these with user-friendly packages, along the lines of Microsoft Office. Behind the interface lies software that converts requests into the formats used by the various existing genome analysis tools, allowing users to screen multiple databases on the web as if they were one. But Andy Baxevanis, director of Computational Genomics at the National Institutes of Health's National Human Genome Research Institute, which is itself about to release a similar system, warns against users becoming too dependent on interfaces as a "black box, where they stick sequences in, get the results out, and don't know what the underlying method actually is". Baxevanis argues that "there is no substitute for actually understanding the individual methods, whether you use them individually or in the context of a workbench".

Collaborating in cyberspace

"By 2010, the relationship of scientists with their computers will change; they will see it as an information and computing fabric that they have customized for the work they are doing, and a window to their colleagues," says Van Houweling. Greater collaboration will emerge naturally as the telephone, video and virtual reality become commonplace on the Internet. Smarr foresees "a persistent cyber café where you walk into a room with your world-wide colleagues."

But there is more to collaboration than sharing a virtual tea-break with colleagues



Computing will be the biologist's number one tool in the future, predicts Craig Venter.

on the other side of the world. A decade ago, scientists studying the upper atmosphere tended to build their research agendas around their own particular instrument, be it a satellite or radar. This hardly made much sense given that the community was studying the same zone. Dan Atkins at the University of Michigan has led an Internet-based effort called the

Upper Atmosphere Research Collaboratory, which has united the community's panoply of instruments and data around common research programmes.

The nature of science is changing, says Lake. "Until now, astronomy has been about getting telescope time, looking at a point in the sky, getting the data, doing something with those data, keeping it proprietary and publishing a small result; but soon all the information is going to be available digitally." Disciplines, he reckons, will increasingly organize large-scale computational resources which are automatically called into play from your PC or workstation. The concept of scientists standing on each others' shoulders is in for a renaissance. Rather than spending most of their time collecting their own data, this will be shared, coded and made accessible to the whole community.

Going for grand challenges

If scientists have access to supercomputing power from their desktops through the Internet by 2010, it will allow them to take on much more ambitious research challenges. "I think we will have the kind of computing power and tools to build models of whole cells and organisms," says Van Houweling, "but these new tools that we are talking about are becoming available only now to ordinary scientists."

Most biologists at present are happy to look for the homologue of the particular gene or protein sequences they are interested in. But by using more powerful computers, it will be possible to start looking at the integration of information across the genome. "If we hope to understand biology, instead of looking at one little protein at a time, which is not how biology works, we will need to understand the integration of thousands of proteins in a dynamically changing environment," says Venter. This is the direction Celera will take next year once it has completed the sequencing of the human and mouse genomes. To do so, the company is building

what will be the world's largest supercomputer for biology, a 1,200-processor machine. "A computer will be the biologist's number one tool," predicts Venter. "The data sets are beyond the capacity of the human brain."

Gene Myers, who leads Celera's computational biology group, has cut the time it takes to piece together the *Haemophilus influenzae* genome from days to just 5 minutes. He is now turning to the task of developing sophisticated algorithms to analyse the human genome, beginning with whole-genome comparisons. He then intends to overlay and integrate gene expression data.

A pointer as to where computational biology is likely to go is a software package called E-cell, developed by Masaru Tomita from Keio University in Japan. The package, which can be downloaded from his website at <http://www.e-cell.org>, simulates basic cellular processes. Tomita has just completed a model of human erythrocytes and is building other models of human mitochondria, signal transduction for chemotaxis in the bacterium *Escherichia coli*, and gene expression networks in this bacterium's lactose operon (a collection of genes that are switched on when the bacterium is forced to feed on lactose sugar). Nourished properly, the "Tamagotchi" erythrocyte reaches a steady state where metabolite concentrations compare well with those reported in real mammalian erythrocytes. Tomita is now inhibiting enzymes from the glycolytic pathway *in silico*, such as hexokinase, glucose-6-phosphate dehydrogenase, phosphofructokinase and pyruvate kinase, in a bid to throw light on the cellular metabolism of people suffering from hereditary anaemia, which is caused by deficiencies of these enzymes. Many are predicting that the fruitfly *Drosophila* will be the main target for complex computer modelling in developmental biology. Its genome will soon be available, and many hands make light work; there are some 6,000 *Drosophila* researchers out there.

There is no shortage of petaflop-scale problems in biology, from brain research to the modelling of organs. The principal limiting step in the development of advanced computing in biology is the lack of biologists who know how to do it. While physicists and astronomers have decades of experience of expressing mathematically complex problems and writing the software to carry them out, computational biology is in its infancy. So physicists will be able to translate more quickly improvements in hardware performance at the petaflop level into the scale or realism of grand challenges. One such challenge is to design a virtual sustained thermonuclear fusion reactor on a computer without having to build real prototypes, a problem estimated at 1 to 10 petaflops.

But even physicists are now running up against the difficulties caused by the sheer complexity of the software needed to model

such systems. US vice-president Al Gore has recently launched a plan for federal agencies to carry out a 'Digital Earth' survey, the stated goal of which is to create "a virtual representation of our planet that enables a person to explore and interact with the vast amounts of natural and cultural information gathered about the Earth". Modelling parts of the Earth such as the ocean or atmosphere is difficult enough; the thought of coupling them all into a working model of Earth is sending software engineers back to their drawing boards. "It makes the Sloan Digital Sky survey look easy," says one scientist (see *Nature* 399, 520; 1999).

Scientists like writing code, but software in the future is likely to be built using higher levels of abstraction. Much current inspiration is coming from *A Pattern Language*, a book written by architect Christopher Alexander in 1977, in which he proposes tackling large problems, like building a city, by building up to them. In this approach, all of the smaller-scale models, such as houses and streets, are solved first. Lake predicts that this concept of looking at the sorts of 'services and interfaces' between components of complex systems may be a key to better algorithm design.

Perhaps the most appealing aspect of distributed power is the prospect of an end to proprietary software, thanks to the input, not least, of researchers who should no doubt be concentrating on their science. The success of the Linux open-source operating system has shown that, when it comes to debugging codes, companies seem no match for the collective IQ of thousands of developers on the web.

But the Internet and computers are easier to invent than they are to predict. The biggest uncertainty is perhaps quantum computing. In 1999, Yasunobu Nakamura's team at the NEC Fundamental Research Laboratory in Tsukuba, Japan, demonstrated electrical control of a quantum data bit (or qubit), a possible first step on the road to the eventual creation of a solid-state quantum computer. Whereas today's computers store information as ones or zeros, quantum computers could use non-classical memory states that incorporate both values simultaneously, resulting in an exponential speed-up in computing time. One small quantum computer would easily outstrip all the computing power in the world combined. Although Dmitri Averin, an expert in quantum computing at the State University of New York, Stony Brook, believes that quantum computing will not become a practical tool within the next ten years, he predicts nonetheless that it will be a much bigger research area than today, and will yield exciting new discoveries in quantum mechanics.

Declan Butler is European correspondent at *Nature*, 3 rue de l'Arrivée, B.P. 264, 75749 Paris Cedex 15, France.

The future of evolutionary developmental biology

Peter W. H. Holland

Combining fields as diverse as comparative embryology, palaeontology, molecular phylogenetics and genome analysis, the new discipline of evolutionary developmental biology aims at explaining how developmental processes and mechanisms become modified during evolution, and how these modifications produce changes in animal morphology and body plans. In the next century this should give us far greater mechanistic insight into how evolution has produced the vast diversity of living organisms, past and present.

Evolutionary biology and developmental biology, or embryology, have had a stormy relationship over the past hundred years. At the end of the nineteenth century and the start of the twentieth, they were inseparable; comparisons of the embryonic development of different species were used as evidence for evolution, while evolutionary history was seen as sufficient explanation for almost every structure or process observed in animal development^{1,2}. But as evolutionary theory embraced genetics in the 1920s and 1930s, the study of embryonic development was largely rejected as being insufficiently precise or quantitative to contribute to an increasingly rigorous science.

The past 15 years have seen a timely reconciliation between the two fields, with the vibrant new discipline of evolutionary developmental biology emerging at the interface. This concerns itself with how developmental processes themselves have evolved: how they can be modified by genetic change, and how such modifications produce the past and present diversity of morphologies and body plans. Three main factors have contributed to the emergence and phenomenal growth of evolutionary developmental biology. Ironically, all three depend on genetics — the discipline that split evolution and development apart 60 years earlier.

Man is but a worm?

The first factor, and arguably the most important, was the discovery that animals as different as nematodes, flies and mammals use similar genes for similar developmental purposes, such as controlling the development of spatial organization in the embryo. The ball started rolling with the discovery in 1984 of a shared DNA sequence motif — the homeobox — in a variety of genes that control development in the fruitfly *Drosophila*. This confirmed that genes with distinct

functions in *Drosophila* development were evolutionarily related, that is, they had all derived from the same ancestral gene in some remote and simpler ancestor³⁻⁵.

Even more dramatic was the demonstration that previously unknown genes in other types of animal, including vertebrates, also possessed the homeobox motif⁵. With a few exceptions, developmental biologists were surprisingly cautious in their initial reaction. Most accepted that the homeobox might prove a useful tag for discovering and cloning interesting developmental genes, but largely shied away from the implied suggestion that the homeobox was highlighting developmental mechanisms shared between organisms as distantly related as fruitflies and humans. Clearly, evolutionary and developmental biology were in no rush to form a partnership.

But by 1989 the evidence for common developmental mechanisms was becoming overwhelming. The two principal complexes of homeobox genes in *Drosophila* (the homeotic genes) were proved to be directly related to four clusters of homeobox genes in mammals, and to share with them a similar physical organization and patterns of gene expression. Not only had all these genes derived from the same cluster of genes present in some remote ancestor, but the fruitfly and mammalian gene clusters, the Hox genes, still had a similar and fundamental role in their respective organisms — to specify the identity of different regions along the head-to-tail axis (the anteroposterior axis) of the body⁵.

Evolutionary biologists call this retention of structure and function 'conservation', and other remarkable examples soon followed.



Figure 1 'Man is but a worm' reproduced from Punch's Almanac for 1882. This cartoon originally symbolized the evolution of man, but is pertinent to the striking conservation of developmental mechanisms and genes between vertebrates and bilaterian invertebrates. Reproduced with permission of Punch Ltd.

The *Pax-6* gene turned out to be implicated in eye development virtually throughout the animal kingdom, and a homeobox gene, *tinman*, is involved in heart development in flies and vertebrates. The discovery of conservation permits previously impossible comparisons of the development of organisms with very different body plans, and has stimulated developmental biologists to consider the evolutionary ancestry of developmental mechanisms, often for the first time. So many examples of conservation have now been found that it is no longer considered surprising, and the cartoon in Fig. 1 is even more appropriate now than it was in the 1880s. We can now state with confidence that most animal phyla possess essentially the same genes, and that some (but not all) of these genes change their developmental roles infrequently in evolution.

Constructing the tree

The second crucial factor was the rise of molecular phylogenetics — the comparison of nucleic acid sequences from different organisms and the construction of evolutionary trees from these data. In 1988, Field *et al.* published a pioneering paper that tackled the vast task of constructing the lines of descent — the phylogeny — of the entire animal kingdom using comparisons of ribosomal RNA sequences⁶. At the time, invertebrate zoologists were beginning to appreciate the extent of convergent evolution (that is, when animals with quite different evolutionary histories have similar features), and were thus casting doubt on traditional anatomy-based phylogenetic schemes⁷. Molecular phylogenetics, on the other hand, seemed to provide an objective method of assessing evolutionary relationships. Since 1988, methods of DNA sequence analysis have been improved, the range of species sampled has increased, some potential sources of artefact removed⁸, and complementary molecular data added⁹. The current molecular-based view of invertebrate relationships certainly lacks resolution, but at least it provides a framework within which comparative developmental data can begin to be interpreted (Fig. 2).

Technical advances in molecular biology provided the third impetus to evolutionary developmental biology. Low-stringency library screening, the polymerase chain reaction and *in situ* hybridization were each invented or refined in the 1980s, facilitating the cloning and analysis of genes in any species, not just the handful of model species traditionally studied.

In the rest of this article, I outline several directions in which I believe there is a real chance of significant advance in the future. This list is certainly not comprehensive, partly because of my own biases, but also because this area of science is renowned for throwing up surprises.

The limits to conservation

Despite the interest of each new example, simply documenting yet more cases of conservation between *Drosophila*, nematodes and human is likely to add little to our understanding of the actual course of evolution. Attention now needs to be diverted to rigorously determining the limits to conservation in each case.

We need, for example, to establish when and in what type of organism the conserved gene or gene function first appeared; to identify the secondary modifications or losses that have occurred; and, if possible, to detect underlying factors that may have constrained or promoted change. The example of the Hox gene clusters (Box 1) illustrates the general problem.

This example also shows the crucial importance of having a sound phylogeny on which to base comparative genetic and developmental analyses. As is evident from Fig. 2, knowledge of phylogeny is vital if apparent structural or functional similarities are to be interpreted safely as true homologies — similarities due to descent from the same ancestral source — and the limits to conservation of each homology are to be defined precisely.

Genotype into phenotype

The link between the genetic make-up of an organism (its genotype) and its form and function (its phenotype) lies at the heart of

evolutionary developmental biology. A fundamental question to be addressed is how alterations to the genotype as a result of mutation are transformed through the intermediary of development into changes in form. There are several parts to this question. The first is the relative importance of mutations that alter the coding sequence of a gene, and thus the structure and function of the protein it encodes, as against regulatory mutations that affect the site, timing or level of gene expression. We suspect that regulatory mutations will be most important, although both categories of mutation have been documented in developmentally important genes. Dissecting their individual contributions to developmental change is complicated by the fact that both types of mutation can occur in the same gene.

The importance of regulatory mutations can be seen within the vertebrates, where the anterior boundaries of expression of particular Hox genes have altered significantly in different animals. These changes correlate closely with anatomical changes along the anteroposterior axis, such as the location of the junction between neck and thorax^{10,11}. Similarly, the areas of expression of Hox genes responsible for specifying thoracic identity are greatly expanded in python embryos compared with other vertebrates, which mirrors the anatomical extension of thoracic identity along most of the vertebral column in snakes¹².

Box 1 How conserved are Hox genes?

One outstanding question is whether clustered Hox genes specifying regional identity along the anteroposterior axis are present in all multicellular animals and, indeed, whether their origin marks the origins of the animals themselves. Hox gene clusters have been found so far in chordates (the phylum to which vertebrates belong), echinoderms (the starfish, sea urchins and their relatives), arthropods (crustacea and insects), nemertean worms and nematodes. If we map this distribution onto the working framework of animal evolution provided by recent molecular phylogenies^{8,9}, we can infer that a Hox gene cluster existed in the last common ancestor of all extant bilaterians (animals with (usually) bilateral symmetry and embryos with three germ layers) and is designated B in Fig. 2. This organism was thus the ancestor of three major groupings of animals — the deuterostomes (for example, chordates and echinoderms), ecdysozoans (for example, arthropods) and lophotrochozoans (for example, molluscs and annelid worms). We can also safely assume that all phyla descended from this common ancestor possess (or possessed) a Hox gene cluster.

We cannot, however, draw the same conclusion for animal lineages that split off earlier than this common ancestor, notably the cnidarians (corals, sea anemones and jellyfish), ctenophores (comb jellies), sponges and placozoans. Genes described as Hox genes on the basis of their DNA sequence have been reported from each of these phyla, but physical clustering has not been demonstrated. Furthermore, the discovery in bilaterians of ancient clustered genes similar in DNA sequence to Hox genes but with different roles implies that it is inappropriate to designate a gene as Hox on the basis of sequence alone²². Hence, we do not yet know if a Hox gene cluster was present in the common ancestor of all animals (designated A in Fig. 2), or indeed even earlier.

We also know relatively little about the extent to which the Hox gene cluster has been secondarily modified in different lineages. There are examples of rapid sequence change in some arthropod Hox genes, cluster breakage in *Drosophila* and nematodes⁹, and loss of some Hox genes in pufferfish²⁸ and possibly barnacles²⁹. Gene loss during evolution is poorly documented but could be common; analysis of developmental genes in parasitic species (which tend to lose functions compared with their non-parasitic relatives) could be useful in this regard. Until the origin of the Hox genes is established and the pattern of their subsequent divergence is clarified, we can no longer be satisfied with statements that simply describe Hox gene clusters as 'conserved'.

The mutations responsible for these alterations have not yet been identified; they could have occurred in the regulatory sequences of the Hox genes themselves, or in other genes that control the activation or stabilization of Hox gene expression. DNA sequence analysis of many more species, chosen with regard to their phylogenetic positions, will be the key to finding candidate mutations, whose effects can then be investigated experimentally. Indeed, such a comparative approach was used recently to identify a mutation in an enhancer of the *Hoxc-8* gene of baleen whales¹³.

How important is gene duplication?

A second question in need of resolution is the importance of gene duplication in the evolution of development. Are there developmental processes that are possible with two copies of a gene but not with one? This is a controversial suggestion, not least because it implies that gene number could impose tight genetic constraints on evolution. But although gene duplications may have been very important in some lineages, as discussed below for the vertebrates, it seems that they are neither necessary nor sufficient for most developmental evolution.

Vertebrates, however, are an interesting case. They have many more Hox genes than invertebrates, as a result of the duplication and reduplication of a single ancestral Hox gene cluster. This was accompanied by an expansion of many other developmentally important gene families¹⁴ and, indeed, the total number of genes in the genome¹⁵ (C in Fig. 2 shows when this occurred). It is tempting to speculate that the origin of the complex vertebrate body plan, with its novel cell types and organ systems, was made possible by the availability of extra genetic raw material, in particular, interacting sets of genes that could be gradually recruited for new developmental roles. This hypothesis predicts a tendency for vertebrate genomes to retain duplicates of developmentally important genes but to lose duplicates of the housekeeping genes that control routine metabolic functions, and predicts that the retained genes would tend to acquire new roles in cell types unique to vertebrates. By testing such predictions, it should be possible to determine whether gene duplication was positively exploited in early vertebrate evolution.

The mechanisms by which duplicate genes acquire new roles also needs resolving. The simplest model suggests that one copy of a duplicate gene retains the ancestral role, while additional copies are free to accumulate mutations and diversify. Current evidence indicates that this is an oversimplification. Members of families of duplicated developmental genes in vertebrates often show overlapping spatial and temporal patterns of expression, and also overlapping functions. This suggests that duplicate genes

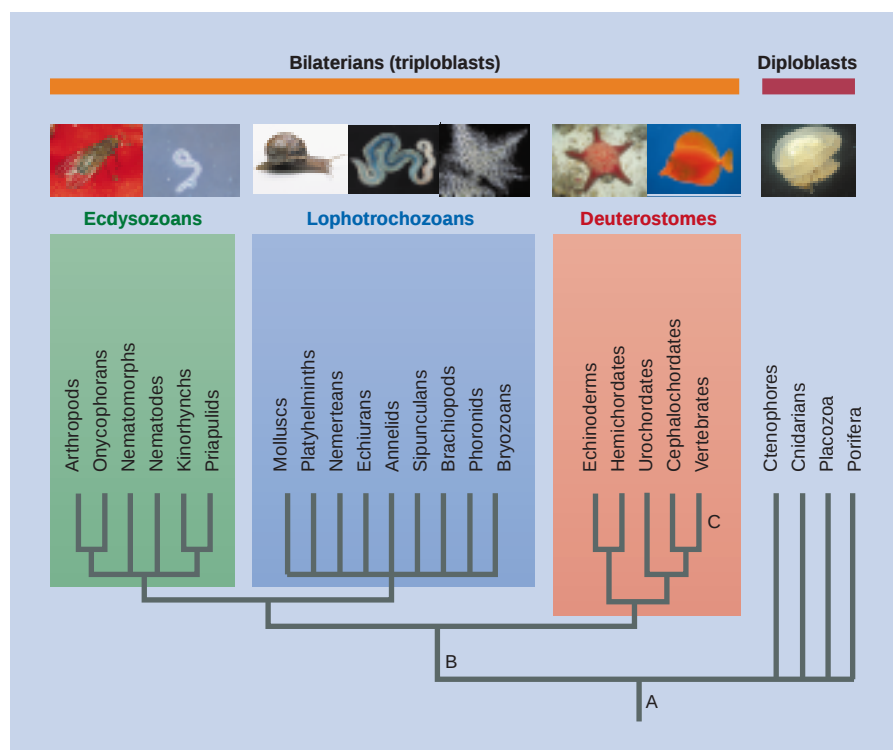


Figure 2 Proposed phylogeny of the animal kingdom, based primarily on 18S rDNA sequences^{6,8} and Hox gene cluster composition⁹. All animals can be divided into two basic groups according to the number of layers in the embryonic body wall: the bilaterians (triploblasts) have three, and the diploblasts have two. The names at the tips of the branches indicate the main animal phyla, taxonomic groupings of animals that share the same body plan. The urochordates, cephalochordates and vertebrates comprise the chordate phylum. Several minor phyla are omitted. More than one branch leading from a split (for example at the base of the ecdysozoans or lophotrochozoans) indicates lack of resolution or lack of consensus given the available data. Branches labelled A, B, C refer to the origin of animals, of bilaterians and of vertebrates, respectively; see text for discussion.

may retain the shared ancestral role, but supplement this with new roles.

A very similar picture of duplicate gene expression could, however, be achieved in a totally different way. If each duplicate gene loses some of its suite of original functions, control of a complex developmental process would become divided between the multiple descendants of the original gene¹⁶. The relative importance of these different modes of evolution needs elucidating, and not only with respect to vertebrates.

A new 'New Synthesis'

In all the scenarios outlined above, the creation of genetic variation is seen as fundamentally important to the rate, timing or pattern of evolution. At first sight, this viewpoint does not sit comfortably with the neodarwinian 'New Synthesis', which places more emphasis on the reduction of genetic variation during evolution as a population becomes increasingly adapted to its environment and the desirable genes that bring this about spread throughout the population. If developmental biology is to be fully integrated with evolutionary biology this conceptual gap needs to be bridged.

Some steps have already been taken. First,

it is now clear that we need not think of mutations in developmental control genes (such as Hox genes) as always causing large phenotypic changes. For example, small differences in trichome patterns on the second leg of different *Drosophila* species seem to be the result of subtle differences in the regulation of the Hox gene *Ubx*¹⁷. We need to determine whether numerous mutations of very small phenotypic effect or fewer mutations of larger effect generally contribute to morphological differences between species¹⁸.

Second, it is now clear that populations can harbour extensive genetic variation with the potential to cause morphological change but which is only revealed under particular conditions. For example, perturbation of the *Drosophila* heat-shock protein Hsp90 by mutation or changes in environmental conditions uncovers phenotypic variation generated by otherwise hidden genetic variation¹⁹.

These findings may point the way towards a logical framework for the 'microevolution' of development — the generation of small genetically determined differences in development that lead to the relatively minor variations in morphology. We can assume that mutations of small phenotypic consequence

arise frequently, and are either exposed immediately to selection pressures or sheltered temporarily in cryptic form by interactions with other genes. Mutations with larger phenotypic effects can also arise, but it is unclear how frequently they contribute to morphological evolution. An extreme view is that the largest-effect mutations serve only to stabilize morphological change already produced by the gradual accumulation of small changes²⁰.

The huge challenge for the future is to convert this conceptual framework into a quantitative model, with parameters such as magnitude of phenotypic effect (and its heterogeneity), number of genes involved, mutation rates, effects of genetic recombination, gene additivity and effects of the environment. It will then be important to examine the model's behaviour in relation to variables such as population size or degree of fragmentation, selection pressure and genetic drift.

It remains to be seen whether the lessons learnt from the study of microevolution will be sufficient to explain the much greater morphological and physiological differences between higher taxa such as phyla. I suspect that radical alterations to genetic systems (for example, duplication of the whole genome or major alterations in the mechanism of gene regulation) will need to be included if we are to explain some truly major transitions, such as the origin of multiple germ layers in the ancestor of the bilaterians (B in Fig. 2) or the emergence of the vertebrates.

Vermes of the Vendian

Fifteen years ago, few developmental biologists would have heard of the Ediacaran fauna, and few palaeontologists would have confessed to an interest in *Drosophila* genetics. Much has changed. Palaeontologists and developmental biologists are now regularly combining data to tackle key questions in evolutionary developmental biology. Consider timescales. Palaeontology is vital to estimating when a particular evolutionary change occurred; it can also reveal whether such a change was correlated with other evolutionary events or with environmental change. The Cambrian explosion of animal phyla is the classic example, but remains controversial. Palaeontology clearly records a rapid increase in the abundance and diversity of animal fossils at the base of the Cambrian, but disagreement abounds as to whether this reflects increases in body size, the origin of shells and skeletons, or a true rapid diversification of body plans. Even if the last explanation is accepted, did this burst of evolution occur in the ancestors of all multicellular animals, or just among the bilaterian lineage (A versus B in Fig. 2)?

Resolving these questions is important, not least because it would indicate where (and whether) to search for possible internal

genetic or external environmental triggers for animal diversification. Several lines of evidence need to be combined to move the debate forward. The continued study of Precambrian (Vendian) fossils is clearly important. If any of these can convincingly be shown to be allied to extant bilaterian phyla, this would cast doubt on the idea of the rapid diversification of bilaterian lineages in the Cambrian. Further analysis of developmental control genes in cnidarians and ctenophores will also be useful, as this might reveal whether bilaterians indeed have unique developmental characteristics^{21,22}.

Another major contribution of palaeontology is in the recognition of ancestral character states and extinct character combinations. These can help us deduce the actual path of developmental evolution, given the genetic and developmental data at our disposal from extant organisms. Our understanding of the early evolution of vertebrates²³, the radiation of the lophotrochozoans²⁴ and the evolution of the tetrapod limb²⁵ have all benefited from such data. Evolutionary developmental biology will continue to benefit from fossil evidence, provided that the science of palaeontology continues to be widely appreciated.

Embracing genomics

The sequencing of complete genomes from multicellular organisms promises to revolutionize the biological sciences. What are the implications for developmental biology? Animal developmental biologists will probably have to be content in the foreseeable future with a nematode or two, a couple of insects, human, mouse and two rather advanced fish. The opportunities will still be huge. The complete genome sequence of the nematode *Caenorhabditis elegans* has already yielded surprises, including some previously undiscovered Hox genes, secondary loss of the Hedgehog signalling molecule and one of its receptor components, and an exceptionally large number of genes for steroid-hormone receptors^{9,26}. As each genome is sequenced, it will yield its own set of lineage-specific expansions and reductions within families of homologous genes. Eventually, a crude but complex picture should emerge of the general pathways of genome diversification during evolution. It will then be a major task to determine which genomic changes relate to modifications in developmental control or morphology. The pattern of differences should, however, give immediate clues as to which sorts of gene families, or genetic pathways, are prone to change, and which may be evolutionarily constrained.

From the perspective of understanding how animal body plans evolve, it is unfortunate that the genome sequences most likely to be completed first do not encompass a particularly wide range of the body plans present within the animal kingdom. Obvi-

ous examples to add include a cnidarian, a lophotrochozoan (for example a mollusc or annelid worm), an echinoderm, and a urochordate (a tunicate) or cephalochordate (amphioxus).

Complete genomes provide more than simply a catalogue of genes. For example, because of the roles of chromatin structure and nuclear architecture in gene regulation²⁷, neighbouring genes could be subject to coordinated regulation. We might then expect the position of a gene on a chromosome to be functionally relevant and conserved in some cases. Once several complete genomes are available the hypothesis of conserved gene position can be tested. It seems quite possible that the correspondence between a Hox gene's position in a gene cluster and its expression along the anteroposterior axis, so fundamental to patterning the bilaterian body plan, may be just the tip of an iceberg.

Peter W. H. Holland is at the School of Animal and Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK. e-mail: p.w.h.holland@reading.ac.uk

1. Darwin, C. *The Origin of Species by Means of Natural Selection* (John Murray, London, 1859).
2. Haeckel, E. *The Evolution of Man: A Popular Exposition of the Principal Points of Human Ontogeny and Phylogeny* (Appleton, New York, 1896).
3. McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A. & Gehring, W. J. *Nature* **308**, 428–433 (1984).
4. Scott, M. P. & Weiner, A. J. *Proc. Natl Acad. Sci. USA* **81**, 4115–4119 (1984).
5. McGinnis, W. *Genetics* **137**, 607–611 (1994).
6. Field, K. G. et al. *Science* **239**, 748–753 (1988).
7. Willmer, P. G. *Invertebrate Relationships: Patterns in Animal Evolution* (Cambridge Univ. Press, 1990).
8. Aguinaldo, M. A. et al. *Nature* **387**, 489–493 (1997).
9. de Rosa, R. et al. *Nature* **399**, 772–776 (1999).
10. Burke, A. C., Nelson, C. E., Morgan, B. A. & Tabin, C. *Development* **121**, 333–346 (1995).
11. Gaunt, S. J., Dean, W., Sang, H. & Burton, R. D. *Mech. Dev.* **82**, 109–118 (1999).
12. Cohn, M. J. & Tickle, C. *Nature* **399**, 474–479 (1999).
13. Shashikant, C. S. et al. *Proc. Natl Acad. Sci. USA* **95**, 15446–15451 (1998).
14. Holland, P. W. H. & García-Fernández, J. *J. Dev. Biol.* **173**, 382–395 (1996).
15. Simmen, M. W., Leitgeb, S., Clark, V. H., Jones, S. J. M. & Bird, A. *Proc. Natl Acad. Sci. USA* **95**, 4437–4440 (1997).
16. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. *Genetics* **151**, 1531–1545 (1999).
17. Stern, D. L. *Nature* **396**, 463–466 (1998).
18. Mackay, T. F. C. *BioEssays* **18**, 113–121 (1996).
19. Rutherford, S. L. & Lindquist, S. *Nature* **396**, 336–342 (1998).
20. Budd, G. E. *BioEssays* **21**, 326–332 (1999).
21. De Robertis, E. M. & Sasai, Y. *Nature* **380**, 37–40 (1996).
22. Brooke, N. M., García-Fernández, J. & Holland, P. W. H. *Nature* **392**, 920–922 (1998).
23. Aldridge, R. J. & Purnell, M. A. *Trends Ecol. Evol.* **11**, 463–468 (1996).
24. Conway Morris, S. & Peel, J. S. *Phil. Trans. R. Soc. Lond. B* **347**, 305–358 (1995).
25. Coates, M. I. *Development* (Suppl.) 169–180 (1994).
26. Ruvkun, G. & Hobert, O. *Science* **282**, 2033–2041 (1998).
27. Cockell, M. & Gasser, S. M. *Curr. Opin. Gen. Dev.* **9**, 199–205 (1999).
28. Aparicio, S. et al. *Nature Genet.* **16**, 79–83 (1997).
29. Mouchel-Vielh, E., Rigolot, C., Gibert, J.-M. & Deutsch, J. S. *Mol. Phylogenet. Evol.* **9**, 382–389 (1998).

Acknowledgements. I thank M. Cohn, S. Shimeld and A. Holland for comments on the manuscript, B. Okamura for the bryozoan photograph in Fig. 2, and B. Cohen, M. Telford and other colleagues for helpful discussions. I hope that non-animal biologists will excuse my zoocentric selection of examples.

Feeding the world in the twenty-first century

Gordon Conway and Gary Toenniessen

The gains in food production provided by the Green Revolution have reached their ceiling while world population continues to rise. To ensure that the world's poorest people do not still go hungry in the twenty-first century, advances in plant biotechnology must be deployed for their benefit by a strong public-sector agricultural research effort.

The Green Revolution was one of the great technological success stories of the second half of the twentieth century. Because of the introduction of scientifically bred, higher-yielding varieties of rice, wheat and maize beginning in the 1960s, overall food production in the developing countries kept pace with population growth, with both more than doubling. The benefits of the Green Revolution reached many of the world's poorest people. Forty years ago there were a billion people in developing countries who did not get enough to eat, equivalent to 50 per cent of the population of these countries. If this proportion had remained unchanged, the hungry would now number over two billion — more than double the current estimate of around 800 million, or around 20 per cent of the present population of the developing world. Since the 1970s, world food prices have declined in real terms by over 70 per cent. Those who benefit most are the poor, who spend the highest proportion of their family income on food.

The Green Revolution brought benefits too for the industrialized world. The high-yielding varieties of staple crop plants bred by the international agricultural research centres of the CGIAR (the Consultative Group on International Agricultural Research) have been incorporated into the modern varieties grown in the United States and Europe. The additional wheat and rice produced in the United States alone from these improved varieties is estimated to have been worth over \$3.4 billion from 1970 to 1993 (ref. 1).

Yet today, despite these demonstrable achievements, over 800 million people consume less than 2,000 calories a day, live a life of permanent or intermittent hunger and are chronically undernourished². Most of the hungry are the women and young children of extremely poor families in developing countries. More than 180 million children under five years of age are severely underweight: that is, they are more than two standard deviations below the standard weight for their age. Seventeen million children under five die each year and malnourishment contributes to at least a third of these deaths.

As well as gross undernourishment, lack of protein, vitamins, minerals and other micronutrients in the diet is also widespread³. About 100 million children under five suffer from vitamin A deficiency, which can lead to eye damage. Half a million children become partly or totally blind each year, and many subsequently die. Recent research has shown that lack of vitamin A has an even more pervasive effect, weakening the protective barriers to infection put up by the skin, the mucous membranes and the immune system⁴. Iron deficiency is also common, leading to about 400 million women of childbearing age (15–49 years) being afflicted by anaemia. As a result they tend to produce stillborn or underweight children and are more likely to die in childbirth. Anaemia has been identified as a contributing factor in over 20 per cent of all maternal deaths after childbirth in Asia and Africa.

If nothing new is done, the number of the poor and hungry will grow. The populations of most developing countries are increasing rapidly and by the year 2020 there will be an additional 1.5 billion mouths to feed, mostly in the developing world. What is the likelihood that they will be fed?

The end of the Green Revolution

The prognosis is not good. As indicated in Fig. 1, there is widespread evidence of decline in the rate of increase of crop yields^{5–7}. This slowdown is due to a combination of causes. On the best lands many farmers are now obtaining yields close to those produced on experimental stations, and there has been little or no increase in the maximum possible yields of rice and maize in recent years. A second factor is the cumulative effect of environmental degradation, partly caused by agriculture itself.

Simply exporting more food from the industrialized countries is not a solution. The world already produces more than enough food to feed everyone if the food were equally distributed, but it is not. Market economies are notoriously ineffective in achieving equitable distribution of benefits. There is no reason to believe that the poor

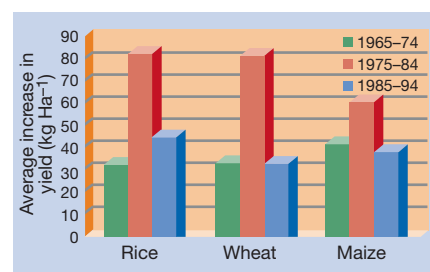


Figure 1 Average annual increase in yields of rice, wheat and maize in developing countries by periods.

who lack access to adequate food today will be any better served by future world markets. Food aid programmes are also no solution, except in cases of specific short-term emergency. They reach only a small portion of those suffering chronic hunger and, if prolonged, create dependency and have a negative impact on local food production.

About 130 million of the poorest 20 per cent of people in developing countries live in cities. For them, access to food means cheap food from any source. But 650 million of the poorest live in rural areas where agriculture is the primary economic activity, and as is the case in much of Africa, many live in regions where agricultural potential is low and natural resources are poor⁸. They are distant from markets and have limited purchasing power. For them, access means local production of food that generates employment and income, and is sufficient and dependable enough to meet local needs throughout the year, including years that are unfavourable for agriculture.

All these arguments point to the need for a second Green Revolution, yet one that does not simply reflect the successes, and mistakes, of the first. In effect, we require a 'Doubly Green Revolution', an agricultural revolution that is both more productive and more 'green' in terms of conserving natural resources and the environment than the first. We believe that this can be achieved by a combination of: ecological approaches to sustainable agriculture; greater participation by farmers in

Table 1 Biotechnology research useful in developing countries

Traits now in greenhouse or field tests	Traits now in laboratory tests
<i>Input traits</i>	<i>Input traits</i>
Resistance to insects, nematodes, viruses, bacteria and fungi in crops such as rice, maize, potato, papaya and sweet potato	Drought and salinity tolerance in cereals
Delayed senescence, dwarfing, reduced shade avoidance and early flowering in rice	Seedling vigour in rice
Tolerance of aluminium, submergence, chilling and freezing in cereals	Enhanced phosphorus and nitrogen uptake in rice and maize
Male sterility/restorer for hybrid seed production in rice, maize, oil-seed rape and wheat	Resistance to the parasitic weed <i>Striga</i> in maize, rice and sorghum, to viruses in cassava and banana, and to bacterial blight in cassava
New plant types for weed control and for increased yield potential in rice	Nematode resistance and resistance to the disease black sigatoka in banana
	Rice with the alternative C ₄ photosynthetic pathway and the ability to carry out nitrogen fixation
<i>Output traits</i>	<i>Output traits</i>
Increased β -carotene in rice and oil-seed rape	Increased β -carotene, delayed post-harvest deterioration and reduced content of toxic cyanides in cassava
Lower phytates in maize and rice to increase bioavailable iron	Increased vitamin E in rice
Modified starch in rice, potato and maize and modified fatty-acid content in oil-seed rape	Apomixis (asexual seed production) in maize, rice, millet and cassava
Increased bioavailable protein, essential amino acids, seed weight and sugar content in maize	Delayed ripening in banana
Lowered lignin content of forage crops	Use of genetically engineered plants such as potato and banana as vehicles for production and delivery of recombinant vaccines to humans
	Improved amino-acid content of forage crops

agricultural analysis, design and research; and the application of modern biotechnology directed towards the needs of the poor in developing countries, which is the subject of the rest of this article.

The price of biotechnology

The application of advances in plant breeding — including tissue culture, marker-aided selection (which uses DNA technology to detect the transmission of a desired gene to a seedling arising from a cross) and genetic engineering — are going to be essential if farmers' yields and yield ceilings are to be raised, excessive pesticide use reduced, the nutrient value of basic foods increased and farmers on less favoured lands provided with varieties better able to tolerate drought, salinity and lack of soil nutrients.

In the industrialized countries the new life-science companies, notably the big six multinationals — Astra-Zeneca, Aventis, Dow, Dupont, Monsanto and Novartis — dominate the application of biotechnology to agriculture. In 1998, 'genetically modified (GM)' crops, more accurately referred to as transgenic or genetically engineered crops, mostly marketed by these companies or their subsidiaries, were grown on nearly 29 million hectares worldwide (excluding China)⁹. That year, 40 per cent of all cotton, 35 per cent of soya beans and 25 per cent of maize grown in the United States were GM varieties.

So far, the great majority of the commercial applications of plant genetic engineering have been for crops with single-gene alterations that confer agronomic benefits such as resistance to pests or to herbicides. These agronomic traits can reduce costs to the farmer by minimizing applications of

insecticides and herbicides. However, as with many agricultural inputs, the benefits received by farmers vary from year to year.

Most of the GM crops currently being grown in developing countries are cash crops; Bt cotton, for example, has reportedly been taken up by over a million farmers in China. But despite claims to be 'feeding the world', the big life-science companies have little interest in poor farmers' food crops, because the returns are too low. National governments, the international research centres of the CGIAR, and a variety of western donors are, and will continue to be, the primary supporters of work that produces advances in biotechnology useful to poor farmers. New forms of public-private collaboration could help to ensure that all farmers and consumers benefit from the genetic revolution and, over time, this should increase the number of farmers who can afford to buy new seeds from the private sector.

The cost of accomplishing this will not be insignificant but it should not be excessive. For example, over the past 15 years, the Rockefeller Foundation has funded some US\$100 million of rice biotechnology research and trained over 400 scientists from Asia, Africa and Latin America. In several places in Asia there is now a critical mass of talented scientists who are applying the new tools of biotechnology to rice improvement. To date, most of the new varieties are the result of tissue culture and marker-aided selection techniques. For example, scientists at the West Africa Rice Development Association have used anther culture to cross the high-yielding Asian rices with traditional African rices. The result is a new plant type that looks like African rice during its early stages of growth (it is able to shade out weeds,

which are the most important constraint on crop production in Africa; Fig. 2a) but becomes more like Asian rice as it reaches maturity, thus giving higher yields with few inputs. Marker-aided selection is being used to breed rice containing two or more genes for resistance to the same pathogen, thereby increasing the durability of the resistance, and to accumulate several different genes contributing to drought tolerance.

Potential of genetic engineering

For some time to come, tissue culture and marker-aided selection are likely to be the most productive uses of biotechnology for cereal breeding. However, progress is being made in the production of transgenic crops for the developing countries. As in the industrialized countries, the focus has been largely on traits for disease and pest resistance, but genes that confer tolerance of high concentrations of aluminium (found in many tropical soils) have been added by Mexican scientists to rice and maize (Fig. 2c), and Indian scientists have added two genes to rice which may help the plant tolerate prolonged submergence. There is also the possibility of increasing yield ceilings, through more efficient photosynthesis, for example, or by improved control of water loss from leaves through regulation of stomatal opening and closing¹⁰.

In addition to generating new traits that enable the plant to grow better (input traits), which are useful to poor farmers, GM technology can also generate plants with improved nutritional features (output traits) of benefit to poor consumers. One of the most exciting developments so far has been the introduction of genes into rice that result in the production of the vitamin A

precursor β -carotene in the rice grain¹¹. β -carotene is a pigment required for photosynthesis and is synthesized in the green tissues of all plants, including rice, but is not usually present in non-photosynthetic tissues such as those of seeds. Traditional plant breeding has given us some plants that produce β -carotene in non-photosynthetic tissue, such as the roots of carrots, but despite decades of searching no rice mutants had been found that produce β -carotene in the grain, so conventional breeding was not an option. To get the cells of the grain to produce β -carotene, genetic engineers added three genes for key enzymes for β -carotene biosynthesis to the rice genome. The grain of the transgenic rice has a light golden-yellow colour (Fig. 2b) and contains sufficient β -carotene to meet human vitamin A requirements from rice alone. This 'golden' rice offers an opportunity to complement vitamin A supplementation programmes, particularly in rural areas that are difficult to reach. These same scientists and others have also added genes to rice that increase the grain's nutritionally available iron content by more than threefold.

Over the next decade we are likely to see much greater progress in multiple gene introductions that focus on output traits or on difficult-to-achieve input characteristics (Table 1).

The potential benefits of plant biotechnology are considerable, but are unlikely to be realized unless seeds are provided free or at nominal cost. This will require heavy public investment by national governments and donors, at times in collaboration with the private sector, both in the research and in the subsequent distribution of seed and technical advice. Breeding programmes will also need to include crops such as cassava, upland rice, African maize, sorghum and millet, which are the food staples and provide employment for the 650 million rural poor who need greater stability and reliability of yield as much as increased yield.

The role of the public sector

None of this will happen through marketing by multinational seed companies, particularly if they decide to deploy gene-protection technologies, commonly referred to as terminator gene technologies, which will mean that farmers cannot save seed from the crop and sow it to get the next crop. In developing countries roughly 1.4 billion farmers still rely on saving seed for their planting materials and many gain access to new varieties through farmer-to-farmer trade. Much of the success of the Green Revolution was due to the true-breeding nature of the higher-yielding rice and wheat varieties.

While terminator technology is clearly designed to prevent rather than encourage such spread of proprietary varieties among poor farmers, some argue that it will do them no harm because they can still use and

replant new varieties from the public sector. But if the companies tie up enabling technologies and DNA sequences of important genes with patents, and then use terminator technologies to control the distribution of

proprietary seed and restrict its use for further breeding, the public sector will be severely constrained in using biotechnology to meet the needs of the poor.

Rather than using the terminator technology to protect their intellectual property in developing countries, it would be better if seed companies focused on producing hybrid seed combined with plant variety protection (PVP) to protect the commercial production of the seed. Hybrid plants do produce viable seed but it is not genetically identical to the original hybrid seed; it may lack some of the desirable characteristics. Hence, there is still an incentive (for example, increased yield) for farmers to purchase hybrid seed for each planting. However, if such purchase is not possible, farmers can still use a portion of their harvest as seed and obtain a reasonable crop. Such recycling of hybrids is not uncommon in developing countries and is an important element of food security. And with PVP, new varieties can be protected while also becoming a resource that both the private and public sectors can use in further breeding for the benefit of all farmers.

Intellectual property rights

Even assuming that terminator technologies are not used, there is cause for concern about the rights of developing countries to use their own genetic resources, the freedom of their plant breeders to use new technologies to develop locally adapted varieties, and the protection of poor farmers from exploitation. In part, these concerns result from the privatization of crop genetic improvement, the rapid expansion of corporate ownership of key technologies and genetic information and materials, and the competitive pressure on these companies to capture world market share as rapidly as possible.

It is only recently that intellectual property rights (IPR) have become an important factor in plant breeding, primarily through the greater use of utility patents. Such patents have stimulated greater investment in crop improvement research in industrialized countries, but they are also creating major problems and potentially significant additional expense for the already financially constrained public-sector breeding programmes that produce seeds for poor farmers.

The success of the Green Revolution was based on international collaboration which included the free exchange of genetic diversity and information. Most of the 'added value' present in modern crops has been accumulated over the centuries by farmers themselves as they selected their best plants as the source of seed for the next planting. These 'land races' have traditionally been provided free of charge by developing countries to the world community. The CGIAR centres add value through selective breeding, and the superior varieties they generate are widely

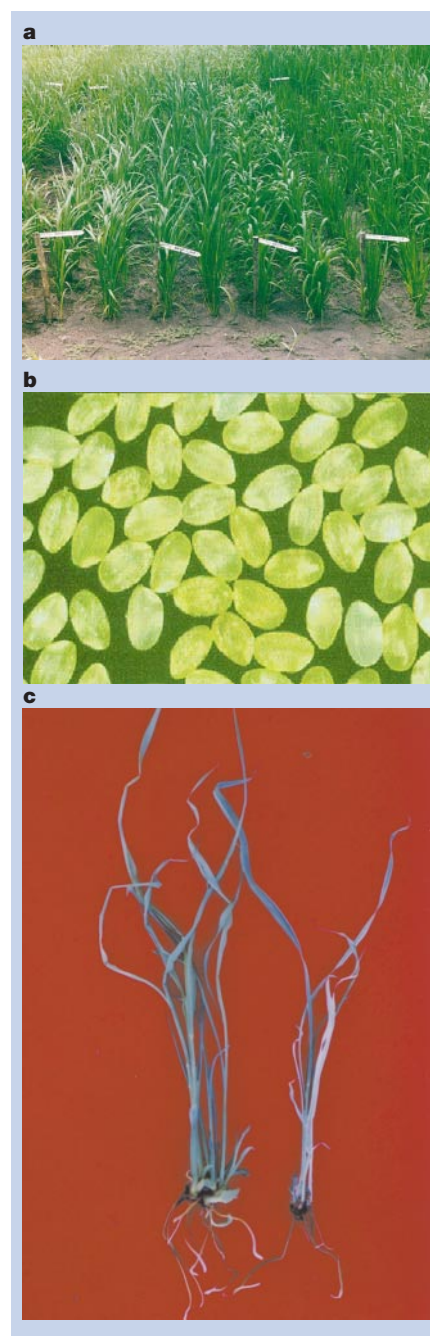


Figure 2 Biotechnology products of value in developing countries. a, Interspecific progenies of Asian $\times$ African rice. Rows 3 and 4 from the right have vigorous growth with droopy lower leaves that suppress weeds. (Courtesy of M. Jones.) b, Transgenic rice grain containing β -carotene (provitamin A). (Courtesy of I. Potrykus and P. Beyer.) c, Transgenic rice (left) with increased citrate production and exudation by roots, tolerates aluminium ($100 \mu\text{M}$ at pH 4.5) better than control. (Courtesy of L. Herrera Estrella.)

distributed without charge, benefiting both developing and developed countries.

Patents on biotechnology methods and materials, and even on plant varieties, are complicating and undermining these collaborative relationships. Public-sector research institutions in industrialized countries no longer fully share new information and technology. Rather, they patent and license and have special offices charged with maximizing their financial return from licensing. Commercial production of any genetically engineered crop variety requires dozens of patents and licenses. It is only the big companies that can afford to put together the IPR portfolios necessary to give them the freedom to operate. And now, under the TRIPS (Trade-Related Aspects of Intellectual Property Rights) agreement of the World Trade Organization, most developing countries are required to put in place their own IPR systems, including IPR for plants. Furthermore, all of this 'ownership' of plant genetic resources is causing developing countries to rethink their policies concerning access to the national biodiversity they control, and new restrictions are likely.

So far, international negotiations relevant to agricultural biotechnology and plant genetic resources have not been effectively coordinated. There are inconsistencies, and the interests of poor farmers in developing countries have not been well represented. The days of unencumbered free exchange of plant genetic materials are no doubt over, and agreements and procedures need to be formulated to ensure that public-sector institutions have access to the technological and genetic resources needed to produce improved crop varieties for farmers in developing countries who will not be well served by the for-profit sector. If the big life-science companies wish to find a receptive and growing market in developing countries, they will need to work with the public sector to make sure this happens.

Some solutions

While negotiations are underway, there are a number of things that should be done. With little competitive loss, seed companies could agree to use the PVP system (including provisions allowing seed saving and sharing by farmers) in developing countries in cooperation with public plant-breeding agencies, rather than using patents or terminator technologies to protect their varieties.

To speed the development of biotechnology capacity in developing countries, companies that have IPR claims over certain key techniques or materials might agree to license these for use in developing countries at no cost.

We would also like to see an agreement to share the financial rewards from IPR claims on crop varieties or crop traits of distinct national origin, such as South Asian Basmati

rice or Thailand's Jasmine rice. The granting of free licenses to use such materials in breeding programmes in the country of origin of the trait might gain the appreciation of developing country researchers and governments.

Finally, the current opposition to GM crops and foods is likely to spread from Europe to the developing countries and maybe even to North America unless there is greater public reassurance. At the heart of the debate about the safety of GM crops and their food derivatives is the issue of relative benefits and risks. The debate is particularly impassioned in Europe. Some of it is motivated by anti-corporate or anti-American sentiment, but underlying the rhetoric are genuine concerns about lack of consumer benefits, about ethics, about the environment and about the potential impact on human health¹²⁻¹⁶.

Much of the opposition tends to lump together the various risks — some real, some imaginary — and to assume there are generic hazards¹⁷. However, GM organisms are not all the same and each provides different potential benefits to different people and different environmental and health risks. Calls for general moratoria are not appropriate. Each new transgene and each new GM crop containing it needs to be considered in its own right. Well planned field tests are crucial, particularly in the developing countries where the risks of using, or not using, a GM crop may be quite different from those in industrialized countries.

The multinational companies could take a number of specific decisions in this area that would improve acceptance of plant biotechnology in both the developing and the industrialized world. First, consumers have a right to choose whether to eat GM foods or not and although there are serious logistic problems in separating crops all the way from field to retail sale, the agricultural seed industry should come out immediately and strongly in favour of labelling. Second, the industry should disavow use of the terminator technology in developing countries and, third, it should phase out the use of antibiotic-resistance genes as a means of selecting transgenic plants. Alternatives exist and should be used.

The Rockefeller Foundation and other donors have invested significant sums in helping developing countries put in place biosafety regulations and the facilities necessary for biosafety testing of new crops and foods, but much more needs to be done. The big life-science companies could join forces and establish a fellowship programme for training developing country scientists in crop biotechnology, biosafety, intellectual property rights and international negotiations administered by a neutral fellowship agency.

Most important of all, a new way of

talking and reaching decisions is required. We believe a global public dialogue is needed which will involve everyone on an equal footing — the seed companies, consumer groups, environmental groups, independent scientists, and representatives of governments, particularly from the developing nations.

Agriculture in the twenty-first century will need to be more productive and less damaging to the environment than agriculture has been in the twentieth. An increased effort is needed to assure that the benefits of agricultural research reach the hundreds of millions of poor farmers who have benefited little from previous research. We believe that biotechnology has significant potential to help meet these objectives but that this potential is threatened by a polarized debate that grows increasingly acrimonious. We need to reverse this trend, to begin working together, to share our various concerns, and to assure the new technologies are applied to agriculture only when this can be done safely and effectively in helping to achieve future food security for our world.

Note added in proof: We commend the Monsanto Company's recent public commitment not to commercialize sterile seed technologies and encourage other companies to follow their lead.

Gordon Conway and Gary Toenniessen are at the Rockefeller Foundation, New York, New York 10018, USA.

1. Pardey, P. G. Alston, J. M., Christian, J. E. & Fan, S. *Summary of a Productive Partnership: The Benefits from U.S. Participation in the CGIAR* (International Food Policy Research Institute, Washington DC, 1996).
2. Conway, G. R. *The Doubly Green Revolution: Food for All in the 21st Century* (Penguin Books, London/Cornell University Press, Ithaca NY, 1999).
3. UNICEF. *The State of the World's Children 1998* (Oxford Univ. Press, Oxford/New York, 1998).
4. Somer, A. & West, K. P. *Vitamin A Deficiency: Health, Survival and Vision* (Oxford Univ. Press, New York and Oxford, 1966).
5. Mann, C. C. *Science* **283**, 310–314 (1999).
6. Cassman, K. G. *Proc. Natl Acad. Sci. USA* **96**, 5952–5959 (1999).
7. Pingali, P. L. & Heisey, P. W. *Cereal Productivity in Developing Countries: Past Trends and Future Prospects*. CIMMYT Economics Paper 99-03 (CIMMYT, Mexico, 1999).
8. Leonard, H. J. in *Environment and the Poor: Development Strategies for a Common Agenda* (ed. Leonard, H. J.) 3–45 (Overseas Development Council, Washington DC, 1989).
9. James, C. *Global Review of Commercialized Transgenic Crops: 1998*. ISAAA Briefs No. 8. (International Service for Acquisition of Agri-biotech Applications, Ithaca NY, 1998).
10. Mann, C. C. *Science* **283**, 314–316 (1999).
11. Ye, X. D. *et al. Science* (submitted).
12. The Royal Society of London. *Genetically Modified Plants for Food Use* (The Royal Society, London, 1998).
13. Nuffield Council on Bioethics. *Genetically Modified Crops: The Ethical and Social Issues* (Nuffield Council on Bioethics, London, 1999).
14. UN Food and Agriculture Organization. *Biotechnology and Food Safety*. FAO Food and Nutrition Paper 61. (World Health Organization/FAO, Rome, 1996).
15. Rissler, J. & Mellon, M. *The Ecological Risks of Engineered Crops* (MIT Press, Cambridge MA/London, 1996).
16. May, R. *Genetically Modified Foods: Facts, Worries, Policies and Public Confidence* (<http://www.2.dti.gov.uk/ost/ostbusiness/gen.html>, 1999).
17. Pretty, J. *The Biochemist* (in the press).

Acknowledgements. We thank M. Lipton, S. Dryden, R. May and colleagues at the Rockefeller Foundation for comments on an earlier draft of this article.

Adapting to climate change

Legislating for the present on the basis of predictions for the future can never be other than a dicey business. The suggestion by a British member of parliament in the nineteenth century that London would be waste-high in horse excrement by the 1950s could have been seen as a call for crippling taxes on Hansom cabs, pushing cabbies out of trade while still leaving us choking on exhaust fumes. ***So should we now be legislating against those exhaust fumes, in fear that by the end of the next century the world will be a flood- and storm-prone hothouse?***

This was one of the key issues highlighted by the Kyoto convention on climate change in 1997. It is all very well to set targets for emissions reductions on the basis of climate models projected over the next century (and then in all probability to fall short of attaining them anyway), but what of the costs of fierce constraints to today's populations, particularly those who live within fragile economies?

Yet clearly we have to do something. There seems to be no serious doubt now that global warming, induced by emission of greenhouse gases through human activity, is upon us. The global average temperature has risen by around half a degree Celsius during the twentieth century, and is predicted to increase by a further 1.0–3.5 °C by 2100. Over the same periods, the seas have risen by 10–25 cm and we can expect almost a metre more at the upper limit. Storms and drought may or may not get more common, but their distribution will probably alter. Massive famine and flooding cannot be ruled out.

As it happens, the Kyoto targets may make very little impact on all of this, but the basic question remains: how can we plan for social change on timescales over which forecasts of enabling technologies become unreliable? One approach that has been advocated is to formulate policy based on assessments of human and social impacts rather than in terms of national emissions reduction targets — something that is harder to quantify but closer to people's concerns. Emissions are not the whole story; what about economic efforts to make societies vulnerable to climate change less so? And how far do political and industrial structures need re-engineering to make actual attainment of objectives realistic?

Climate modellers, meanwhile, are trying to incorporate some element of the social and economic contexts into their forecasts. They grapple with the question of whether it is better, given the uncertainties, to brake hard on emissions now, do nothing until later on the assumption that we will then be much more technologically capable of it (owing to enhanced energy efficiency and so forth), or something in between. Inevitably, such modelling is painfully contentious and strongly influenced by the assumptions that go into it. We are juggling with the unknown; and in the face of massive inertia and sharp political lobbying, that is a precarious occupation.

Philip Ball

The shape of the cosmos

Olaf Stapledon is the great unsung hero of twentieth-century science fiction. His stories are cast on scales of time and space so vast as to inspire nothing short of vertigo or terror. In *Star Maker*, his hero, in a 'hawk-flight of the imagination', journeys through the cosmos and eventually meets the Creator in his workshop. The Star Maker, it seems, is still perfecting his art. Our own cosmos is only provisional. Previous essays in Universe construction litter the room like trash. These essays come in all shapes and sizes, suggesting that we need not live in a great, expanding hypersphere, infinite in all directions.

Our conception of the shape and size of the Universe is based on timescales. The familiar 'light-year' is a unit based on the time it takes for a beam of light to get from A to B. But there is a reckoning that is purely geometric, or, rather, trigonometric. This is the 'parsec', short for 'parallax arcsecond', the distance at which the separation of the Earth and Sun (1 Astronomical Unit) would subtend an angle of one second of arc. A parsec is equivalent to just over three light-years.

But the dimensions of the cosmos are reckoned in megaparsecs — millions of parsecs. ***At the cosmic scale, the uncertainty over the rate at which the Universe is expanding means that time, distance and geometry are inextricably bound together:*** the Hubble flow is meted out in units of kilometres per second per megaparsec. On the very largest scales, time either loses all meaning or becomes subservient to geometry. Our understanding of time may depend on knowing the shape of the cosmos — the shape on the Star Maker's workbench.

In 2002, we may know. By then, we should have results from detailed maps of fluctuations in the cosmic microwave background (CMB) from two spacecraft: NASA's *Microwave Anisotropy Probe* and the European Space Agency's *Planck Surveyor*. The results will help constrain values for important cosmological parameters such as the density of matter in the Universe, and the Hubble constant that provides a scale to its expansion. Accurate knowledge of these parameters should, in turn, constrain ideas about the geometry of space — whether it is 'flat' or 'Euclidean', or if it has positive (that is, spherical) or negative (hyperbolic) curvature.

The results will also be able to tell us whether space is infinite in all directions, or is geometrically finite.

If it is finite, then some extremely weird things could fall out of the analysis. In a finite Universe, surveys of deep space could be pulling out multiple images of the same finite set of galaxies — like an observer in a hall of mirrors. One of those faraway galaxies could be our own, its light having circumnavigated the cosmos to reach us. Study of the CMB will be able to tell us whether space really is a hall of mirrors, and, if it is, something of the hall's architecture. The Universe could be an analogue of a simple sphere. On the other hand, it could have some more exotic shape, akin to a torus, or another of those shapes littering the Star Maker's garage floor.

Henry Gee

Plus ça change

J. L. Heilbron and W. F. Bynum

For the past couple of centuries the penchant for prediction has been prevalent at century turns. How much have evaluations of scientific discovery and predictions for future advancement changed since those of the science commentators at the end of the last century?

Century turns have only recently become occasions for stocktaking, millennially speaking. Perhaps the earliest sustained retrospective occurred around 1800. Among the serious items then produced were histories of science sponsored by the University of Göttingen; their seriousness may be indicated by the length of the entry for physics, which ran to 8,000 pages. The briefer British offered a crisp choice between an improving and a degenerating world, William Godwin (*Essay on Political Justice*, 1794) plumping for progress, Thomas Malthus (*Essay on Population*, 1798) arguing the easier case for decay. But it was not until the turn of the last century that reviewers and previewers created the genre to which this special supplement to *Nature* is devoted. As our contribution to it, we illustrate the proposition that the creators of the genre may have exhausted it.

It is said that ours has been the century of science. That is exactly what the mathematician Maurice Lévy, president of the Paris Académie des Sciences, told his confrères at their last meeting in the last century. (The meeting occurred in December 1900, which shows that the French accepted the opinion aired the preceding January in *Nature*, that centuries begin with years ending in 01.) Lévy observed that only during his lifetime had science descended from the heavens (he had in mind celestial mechanics) to inspire and direct 'industrial mechanics', the application of science to the arts. The results had been astonishing, he said, particularly in the movement of goods, people and information. Would progress continue unabated? Perhaps. Lévy doubted that even heavenly mechanics could solve the next problem in commerce and travel, namely, heavier-than-air flight. But as he rightly observed, scientists should not cast horoscopes.

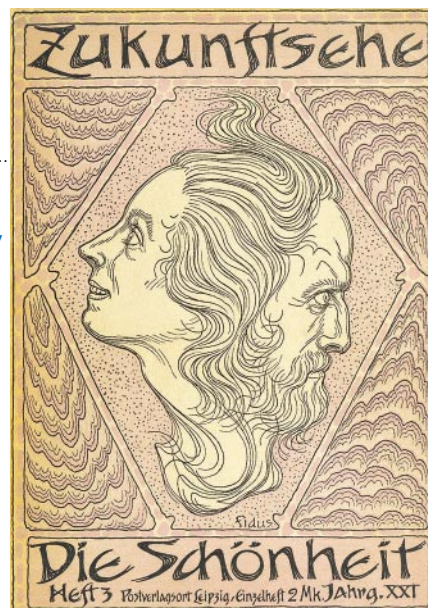
Today we worry that our progress might slacken if emphasis on application dries up funds for research in pure or basic or non-mission-orientated science. Lévy worried about it too. He warned our century against neglecting the essential source of useful applications, that is, useless research. The then editor of *Nature*, Sir Norman Lockyer, took Lévy's warning as the text for his first

sermon of the century. Like Lévy, Lockyer did not doubt that science-based progress would continue; but unlike Lévy, he fretted that it would not take place in Britain: Unless we put more science everywhere, he said, in our schools, industry and government, we will certainly lose out, in accordance with our countryman's doctrine of the survival of the fittest, to Europe and the United States. This cry raised a chorus then as it does now. The chemist Sir Henry Roscoe advised that unless the government poured money into education, "our children and grand children may see England sink to the level of a third rate Power." And a writer in the *Fortnightly* questioned whether "England [would] survive the century."

We continue our theme that every divination or evaluation, optimistic or pessimistic, made about science around 1900 has an analogue in writings dating from our times, under the rubrics science wars, the art of war, the war on disease, successes and failures, new technologies, and new sciences.

Science wars

Recently, especially in the United States, scientists have waged what they began as a defensive war against the carping of humanists of the Luddite and constructivist schools. The defensive scientists objected to the accusation that science produces only pollution, weaponry and other instruments of social control, and, moreover, that it has no claim to truth; like anything else, according to the constructivists, it is the product of negotiation, authority and influence. The law of gravitation is as much social convention as natural necessity. And, say the Luddites, science undermines religion and morality by insisting on evolution and providing the contraceptive pill. Fearing that the humanists' challenge might damage the reputation of science among the general public who pay the research bill, scientists attacked the constructivists' constructions as opaque drivel, as in some cases it is. As for organized religion, where the stakes are higher, the response is more nuanced. There is tolerance and even attempts at rapprochement between some scientists and some theologians, but summary rejection of the beliefs of the creationists.

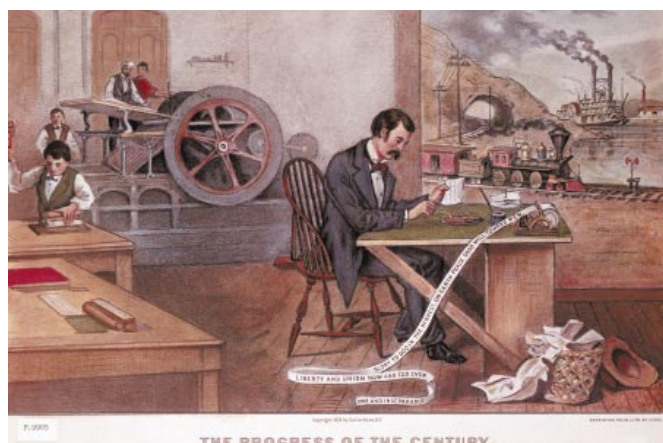


Century ends encourage scientists to look back and forward, like the two-faced Roman god Janus.

We find the same charges and defences at the end of the last century, although admittedly in a livelier literary form. Science is immoral ("[a scientist] will view his mother's tears not as expressions of her sorrow, but as solutions of muriates and carbonates of soda"); anti-religious ("what does it matter whether the conceptions men form of the universe are those of the nineteenth century or of the ninth, in comparison with their acceptance of [religious] truths?"); and arrogant, selfish and dehumanizing ("bring up a woman in the positivist school and you make of her a monster, the very type of ruthless cynicism"). In a famous episode in 1895, a French journalist returned from a visit to the Vatican with the news that science was bankrupt. French scientists responded with a huge banquet in honour of the chemist and minister Marcellin Berthelot, at which speakers took turns bashing the religious right and defenders of the old ways of life. The guardian of morality is not religion, but science, they said, for science inculcates probity, accuracy, tolerance and modesty(!). Science emancipates the human spirit — "humanity knows no other guide." This bombast, widely repeated at the time, has reappeared in almost the same words in our own science wars.

The art of war

Many people condemn the entanglement of science with the military. Others recommend it as a good route to high-tech consumer goods and support of basic research. Much can be said for the latter point of view. The world wars vastly accelerated the development of aviation, radio, electronics and nuclear power; the Cold War



The printing press, telegraph, railroad and steam engine are depicted in a late-nineteenth century lithograph 'The progress of the century'...



...while in 1999, visitors at an industrial fair view the future of engineering at a virtual reality exhibit.

brought the computer; the Race for Space, the teflon pan. All the basic sciences prospered under military funding in the United States and the Soviet Union. And how was it in 1900? Lévy extolled the contributions of the military to nineteenth-century civilization: "Today's cannon is one of the most instructive laboratories of science." Experiments done in gun barrels showed how to contain high pressures in steam engines and internal combustion machines; civilization gained ocean-going steamships, powerful locomotives and better steels for pistons and rails. *Nature* regularly reported on new armaments as advances in applied science.

The pronouncements of the Kaisers are a good quarry for a symbol of the symbiosis of science and the military. In 1900, dressed in the uniform of an officer of engineers and surrounded by generals and admirals, Wilhelm II told an assembly at the Technische Hochschule in Berlin that their brand of science applied to commercial and military warfare was true science, and worthy of the PhD; whereupon he awarded them the right to grant the doctorate in philosophy.

Just as in our time many people believe that atomic bombs prevent world wars, so, at the end of the last century, observers of the art of war, for example Sir Michael Foster and Alfred Nobel, expected that the constantly improving mechanization of warfare in their time would force governments to keep the peace.

The war on disease

The worldwide eradication of smallpox has encouraged optimists to believe that many other diseases will soon go extinct. Polio might be a candidate, and the distribution of malaria, we are told, is currently being rolled back, even while we await the long-promised vaccine. HIV, lassa fever and drug-resistant tuberculosis have tempered current predictions about the conquest of microbial disease, and prions have extended notions of the 'germ'. The lamentable decline in

biodiversity is less obvious at the disease-causing level. We worry, too, about escalating health costs and distribution of resources, about whether doctors have too much or too little power and whether 'managed care' is the answer. In 1900, health costs were also on the rise and many charity hospitals round the world struggled to make ends meet. Several European countries had already introduced national health insurance schemes, Britain was soon to follow and even the United States was flirting with what has since become a heretical idea. Polio was a rare disease, almost as scarce as it has become in the developed world; malaria was already earmarked as a preventable disease; and, even without antibiotics, tuberculosis was perceived as in retreat. Sir William Osler predicted that, by 2000, doctors from the Old World might even journey to the New, in search of inspiration and training. *The Lancet* was convinced that the most striking achievement of nineteenth-century medicine was to abandon the notion of specific drugs against disease; its leader-writer for 5 January 1901 declined to predict how much more would be achieved for practical medicine by science.

Successes and failures

Alfred Russel Wallace was not a man to let so momentous an occasion as the passing of a century go by without summing its pluses and minuses. Co-discoverer of the principle of natural selection, polymath, and a self-made man of integrity, Wallace reckoned that in his "Wonderful Century" scientific, medical and technological successes preponderated over missed opportunities and social disintegration. Like many scientific commentators, he decried the notion of progress in art and design. In science, however, he counted 24 major achievements in the century (including railways, the telephone, the theory of evolution and lucifer matches) against only 15 in all preceding ones.

The century's failures stimulated him to prophesy. There had been a woeful neglect of

phrenology and of psychical research, he insisted, and the delusion that smallpox vaccination could prevent smallpox persisted. In each case, Wallace was convinced that the future would vindicate his position. It has in part: phrenology now has its own website, psychical research has its Koestler chair, and the view that smallpox vaccination played a part in the origin and spread of AIDS has its supporters. But the most important failures of Wallace's century were moral, political or economic: the greedy exploitation of the poor by the rich; the growth in the number of millionaires; the neglect of India and other poor countries; militarism; the desecration of the natural world; the self-serving of corrupt politicians; the failure to realize the vast possibilities for human welfare that science offers. Here, unfortunately, Wallace could have been speaking of our age as well as his own. Whether his optimistic forecast (that "the flowing tide is with us") also applies to us we must wait to see.

New technologies

We take the hardest case. The computer is touted for creating a revolution in the acquisition and processing of information. Profound social and cultural consequences have already followed, for example, in financial transactions and the decentralization of the work place. And, if half the predicted developments are realized in twice the time foreseen, there will be no new books, no shops, no banks — only the computer and the telephone — by the middle of the new century. Without pretending to evaluate degrees of revolution, we must insist that social commentators around 1900 expressed themselves as hyperbolically about the Victorian revolution in communication and transportation as the nerds of our time write about computers. Take Sir William Hunter, speaking as president of the British Association for the Advancement of Science in 1900: "nineteenth-century man has climaxed his advance from barbarism to civilization by

Curious current notions about the effect of gravity in the Southern Hemisphere are matched by nineteenth century geographical oversight. The transatlantic cable speeded communication between Britain and the United States at the end of the nineteenth century (left), while today worldwide communication by e-mail is the norm (right).



Worldwide e-mail

FREE Internet trial

From the Arctic Circle to Australia, you can e-mail anyone, anywhere in the world for the price of a local call. With Claranet it's easy to get online, and our 24 hr helpline is there to answer your queries day or night, 365 days a year.

ClaraNET

0800 358 28 28

CORBIS/THE ADVERTISING ARCHIVES

overcoming space and time...he has studied the ocean with steamships, girdled the earth with electric wire, tunnelled the lofty Alps, spanned the Forth," and so on. The train, steamship and telegraph not only worked the obvious transformations in commerce, communication and conquest, but also changed the practice of science. Suddenly the world knew technical advisory committees, scientific missions, international scientific organizations, international centres for data collection and — that essential ingredient of the modern life of science — international scientific meetings.

Just as cassettes in our time decry the invasion of privacy and the annihilation of society threatened by the computer, so critics in 1900 castigated the new means of transportation and communication as subversive of civilized life. The tram, train and underground pushed cities with their moral and physical pollution into the clean countryside. Max Nordau, a physician who specialized in degeneration, identified travel, information and the telephone as causes of the morbid anxiety from which his patients suffered. Unless, he said, we can adapt to reading acres of newspapers a day, communicating simultaneously by telephone with friends in five continents, and flying around the world in a few hours (Nordau here was more optimistic than Lévy), it is all up with us.

New sciences

When the US Congress cancelled the Super Collider, the search for the material basis of things was forced to stop before the next lowest level, where theories of everything could be tested. Or so wrote the disappointed high-energy physicists. It may not have been mere penury and the end of the Cold War that broke faith with the Greeks; science may have outgrown the quest. Recruitment into particle physics and other physical sciences became more difficult. The biological sciences and the

other disciplines whose futures are described so favourably elsewhere in this issue have superseded what used to be respected as the most fundamental of the sciences.

The conscientious academic advisor today would have to give an intending scientist the same advice that Max Planck received over 100 years ago: opportunities in fundamental physics are restricted, do something else. The advice given to Planck referred to employment, not to substance; fundamental physics then was at an exciting stage, experiencing one Grand Synthesis after another: heat with mechanics, light with heat, electricity with magnetism, all with all. It appeared that basic understanding of the material universe was in place, or almost so, apart from those troublesome clouds that Lord Kelvin spied in the sunny sky of physics. He saw well, since it took relativity to disperse one of the clouds and quantum theory to drive off the other.

The depth of pessimism about the fate of physics came in 1895. The Deutsche Physikalische Gesellschaft was then celebrating its fiftieth anniversary. Its president, worried by the apparent senescence of his science and the recent deaths of three of its greatest representatives, Helmholtz, Hertz and Kundt, could manage only the lukewarm hope that physics would not decline. By the time he came to publish his remarks, he had learned about the discovery of X-rays. Had he known about Röntgen's discovery, he said, "[I] would have concluded in an entirely different tone...expressing my joy that the second fifty years in the life of the society had begun as gloriously as the first."

This flip-flop indicates an important characteristic of scientific prophecy. The prophets typically base their estimates on their experience of the recent past. The future will be bleak or rosy just as the present is. More reliable predictions may, perhaps, be generated from a larger view. Merely to

show the strength of the method — not because we have any faith in the predictions — we shall make a few.

- Because, despite many better reasons to come to an end, civilization has not done so, we predict that Y2K will not kill it.
- Because, despite increasingly noisy threats from computer enthusiasts, the number of new books published in print continues to grow, we predict that the computer will not slay the printed book, at least not before nature slays us.
- *Nature*, however, will not die.
- No matter how little or great the progress made in any science, its practitioners will continue to call for more research in it.
- Experimental high-energy physics may not survive the century.
- Everything we say about the prospects of science in the twenty-first century will have a precise analogue in writings that will appear in print 100 years from now.

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK. W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.



AKG

The future of public health

Barry R. Bloom

Public health deals with the health and well-being of the population as a whole and its achievements over the past century, especially in the richer countries, have been truly impressive. What direction should public health take in the future?

It is an astonishing fact that half of all the increases in life expectancy in recorded history have occurred within this century and that most occurred in the first half of the century, before the introduction of modern drugs and vaccines. Life expectancy in the United States, for example, has risen from 47 years in 1900 to 78 years in 1995. It is interesting to analyse changes in life expectancy across different countries in this century as a function of *per capita* income (Fig. 1). In addition to the obvious increases in life expectancy globally over time, two additional points are worth mentioning.

First, when people are poor, they die young, and miniscule increases in *per capita* incomes can have a major impact on life expectancy. Second, even if you were enormously wealthy in 1900, there were 25 years of life expectancy you could not buy, which in 1990 could be gained even at relatively modest incomes. That which could not be bought in 1900 is, I believe, knowledge of public health in the broad sense. The major gains in health in this century have been attributable largely to the impact of public health and disease prevention, rather than to medical interventions.

Advances in public health

Public health is best distinguished from clinical medicine by its emphasis on preventing disease rather than curing it, and its focus on populations and communities rather than the individual patient. Perhaps because of this emphasis on large numbers of people, the achievements of public health in this century are truly impressive. In the industrialized world we take clean water for granted. It is only the occasional lapses in water quality, such as the outbreak of cryptosporidiosis (caused by the protozoan *Cryptosporidium*) which sickened 440,000 people in Milwaukee in 1993, that remind us how important safe water is to our collective health. In much of the developing world, however, simply drinking water is a high-risk behaviour.

In the past two decades, deaths from heart attacks and stroke in the United States have dropped by 30–50 per cent¹, in part by behaviour changes, in part by primary prevention with medications. Smoking, which is estimated to be responsible for about 20 per cent of all deaths in the United States², has declined from 42 per cent to 25 per cent of adults over 30 years in the United States

(although it has increased ominously to 36 per cent of US teenagers and is still rising). Another contribution to cancer prevention on a large scale was made when tamoxifen, a drug used to treat cancer, was found to reduce the incidence of breast cancer by 45 per cent in women at high risk. This is an example of a drug thought of as a therapeutic in clinical medicine becoming a public health tool in 'primary prevention', that is, preventing disease in individuals already known to be at risk.

Whereas population growth in the industrialized world has reached almost steady-state levels, infant mortality in the United States has declined by 26 per cent in the past decade and is at the lowest levels ever. And even these impressive figures leave the United States ranked at only twenty-fifth worldwide.

Perhaps most dramatic of all is the impact of immunization. Vaccines have eliminated smallpox from the world and polio from the Northern Hemisphere, and have reduced measles, rubella, tetanus, diphtheria and meningitis in many countries to a handful of cases each year, at a saving of millions of lives and billions of dollars. Vaccines remain the most cost-effective intervention known for preventing death and disease. Indeed, such is the success of immunization that this year, for the first time, infectious diseases are no longer the largest cause of death worldwide³.

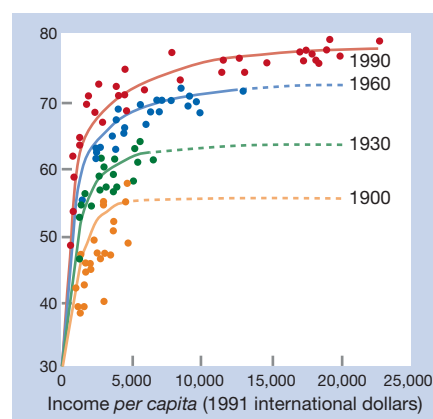


Figure 1 Life expectancy and income *per capita* for selected countries and periods. Figure adapted from ref. 6. The income data, expressed in international dollars, are an estimation of purchasing power and derive from data in ref. 7 and World Bank data.

Disparities remain

The bad news is that the benefits of biomedical science and public health have not been made available to everyone. The disparities are striking. The country with the highest life expectancy is Japan, where people live on average to the age of 80 years; Sierra Leone has the lowest, just 37 years in 1998. And the disparities are not just between countries. Most people believe that in industrialized countries everyone can expect a relatively long life and that the major health issues centre on the quality of life and health care rather than the quantity of life. It is a shock, therefore, to learn that people born in particular rural counties of Minnesota, Colorado, Iowa or Wisconsin on average will live 25 years longer than those born in four counties in South Dakota, 23 years longer than in 12 counties in Mississippi and Alabama, and 22 years longer than people born in Washington, DC, or Baltimore, Maryland⁴. The variance in life expectancy in the United States between women of Japanese extraction in Bergen County, New Jersey, and Bennett County, South Dakota, is 41 years. Overall, disparities in life expectancy between different parts of the United States are greater than for any other nation in the world.

Curiously, the reasons underlying the disparities in the United States are by no means clear, as in many counties the correlation between *per capita* incomes and life expectancy is not particularly good. For example, *per capita* income is significantly higher in the county of Washington, DC, than in several counties along the Texas–Mexico border, yet life expectancy is significantly lower, by about 15 years, in Washington. Understanding and then reducing disparities in health and life expectancy, within and between countries, has to be an important issue on the public health agenda for the next century.

Directions for public health policy

The major thrust of epidemiology in the twentieth century was to analyse the risk factors that contribute to illness and disease. This effort was so successful that we probably now know all the major risk factors. But this is just the beginning. There are four areas in which I believe public health will have an increasingly important role.

First, we need to develop more sensitive



Figure 2 Traffic accidents are becoming a major killer in developing countries.

epidemiological approaches that can identify the many additional risk factors of smaller effect. Also, we need to design better randomized trials so that the biases that prevent us from establishing causation can be eliminated. Second, epidemiological surveillance is needed not only to trace emerging infections but, more generally, systematically to ascertain the burden of disease (that is, the years of healthy life lost because of each disease). Together with analysis of the economic costs of those diseases to individuals and society, and the costs of the interventions available to prevent or treat them, this will enable us to set rational priorities for public spending on research. Such research should be aimed, for example, at establishing where effective interventions are lacking or too expensive to be widely used, and where resource allocations for treatment can result in the most years of healthy life per unit of health expenditure.

While an analysis of cost-effectiveness will be helpful, it will have to go hand in hand with an assessment of the quality of health care and health-care systems in order to set rational priorities in health spending. In the United States alone, a trillion dollars is spent on health care — half of the entire global expenditure on health. Thus, the third area in which public health should have a growing role is in understanding the burdens and costs of interventions, and improving the quality and efficiency of the health service.

Finally, one of the myths of the modern world is that health is determined largely by individual choice and is therefore a matter of individual responsibility. In fact, most behaviour is socially patterned and reinforced in groups. Just providing health information to people is not an effective way to change their behaviour. To make a difference in conditions that we know can be prevented (such as those due to smoking, alcohol and drug abuse, and obesity, hypertension and sexually transmitted diseases) it is essential that we develop a public health approach that will protect populations and establish prevention strategies for groups, not just for individuals.

There are, in fact, some strategies that have made a difference at the population level, with the saving of millions of lives. Local and national decisions have been made to protect the public's health by controlling the environments that cause disease: laws and standards have been adopted that protect us from health hazards in water, food, milk, alcohol and the workplace, and public policies on vaccination have been developed. Now we require a greater understanding of human behaviour, and must learn to develop more effective strategies for changing harmful behaviour. Whereas cigarettes are currently advertised in every country of the world, no one is spending billions of dollars to advertise the pleasures of salmonella or cryptosporidium contamination of the water! The challenge is to change behaviour in the face of massive advertising of tobacco, unhealthy foods and alcohol (which do indeed impart pleasure, at least in the short term). As many health-damaging behaviours are cumulative, we need to reduce unhealthy behaviour in as many people as possible. If we focus at the population level, we will benefit vast numbers of individuals.

Whether public health will have the same impact on health in the twenty-first century as it has in the twentieth is unclear. However, it is not unreasonable to speculate that the unfolding knowledge of the human genome, together with epidemiology and biostatistics (core disciplines in public health) and computational biology, will have an increasingly important role in relating genetic variation and differences in gene expression to individual disease susceptibilities. Studies of large cohorts, such as the Harvard Nurses Study — a group of 121,000 healthy women followed for 23 years⁵ — will enable investigators prospectively to link not only extrinsic and environmental influences such as diet or smoking to the risk of developing a particular disease, but to associate extrinsic risks with intrinsic genetic susceptibility and resistance. This will essentially allow those in the rich countries to create individually tailored drug regimens and behavioural modi-

fications for overcoming predicted individual disease risks, creating a new field of 'boutique medicine'.

But we live in a world in which many countries are not sharing in the benefits of globalization and are not experiencing increased standards of health and quality of life. Much of the knowledge about individual risks that will derive from the Human Genome Project and modern biomedical science, and the resources to obtain the boutique treatments and preventions to overcome these risks, will simply not be available to the 85 per cent of the world's population who comprise the Third World. One can only hope that from the scientific knowledge gained in our current pursuit of sophisticated approaches to treat disease at the individual level, there will also emerge effective, safe and affordable preventions and treatments of relevance at the population level. It is principally these interventions that will make a difference in reducing the global disparities in health and in improving the life expectancy and quality of life of people in even the poorest countries and parts of countries.

We know that cardiovascular disease, infectious diseases, psychiatric disease and physical injuries represent the major global burdens of disease and disability in industrialized and developing countries alike. The challenge for biomedical science and public health in the coming century is to develop the population-based interventions needed to reduce these burdens. Vaccines to prevent AIDS, malaria, tuberculosis, dysentery and other respiratory and diarrhoeal diseases are needed. In addition, effective drugs to prevent as well as to treat cardiovascular disease and psychiatric illness are required, as are effective interventions to prevent injuries, for example, improving road and automobile safety worldwide (Fig. 2) and preventing injuries due to falls in the elderly. That, in my view, will be the major global challenge for biomedical science and public health in the twenty-first century.

Barry R. Bloom is at the Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA.

e-mail: bbloom@hsph.harvard.edu

1. Center for Disease Control. *Deaths: Final Data for 1997* Vol. 47 (National Center for Health Statistics, 1999).
2. McGinnis, J. M. & Foege, W. H. *J. Am. Med. Assoc.* **270**, 2207–2212 (1993).
3. World Health Organization (WHO). *World Health Report 1999* 1–121 (WHO, Geneva, 1999).
4. Murray, C. J. L., Michaud, C. M., McKenna, M. T. & Marks, J. S. *US Patterns of Mortality by County and Race: 1965–1994* 1–97 (Harvard School of Public Health, Cambridge, Massachusetts, and Centers for Disease Control and Prevention, Atlanta, Georgia, 1998).
5. Colditz, G. A. *et al.* The Nurses' Health Study: a 20 year contribution to the understanding of health among women. *J. Womens Health* **6**, 49–62, 1997.
6. World Bank. *World Development Report 1993. Investing in Health* p. 34 (Oxford Univ. Press, 1993).
7. Preston, S. H., Keyfitz, N. & Schoen, R. *Causes of Death: Life Tables for National Populations* (Seminar Press, New York, 1972).

Transitions still to be made

Philip Ball

A collection of many particles all interacting according to simple, local rules can show behaviour that is anything but simple or predictable. Yet such systems constitute most of the tangible Universe, and the theories that describe them continue to represent one of the most useful contributions of physics.

Physics in the twentieth century will probably be remembered for quantum mechanics, relativity and the Standard Model of particle physics. Yet the conceptual framework within which most physicists operate is not necessarily defined by the first of these and makes reference only rarely to the second two. The advances that have taken place in cosmology, high-energy physics and quantum theory are distinguished in being important not only scientifically but also philosophically, and surely that is why they have impinged so forcefully on the consciousness of our culture.

But the central scaffold of modern physics is a less familiar construction — one that does not bear directly on the grand questions that physicists are popularly expected to address but instead defines our current understanding of phenomena at the prosaic energy and length scales characteristic of our everyday experience. Statistical physics, and more specifically the theory of transitions between states of matter, more or less defines what we know about ‘everyday’ matter and its transformations.

Moreover, it provides the conceptual apparatus for tackling complex collective quantum phenomena of intense topical interest such as Bose–Einstein condensation (in which a collection of particles all occupy the same quantum ground state) and high-temperature superconductivity (that is, superconductivity above about 35 K). Many of the states of condensed matter that promise new technological applications, ranging from block copolymers to magnetic multilayers, can be understood as the consequence of the kind of collective behaviour that statistical physics describes.

There are still central issues in cosmology and high-energy physics whose solution requires an understanding of phase transitions, not least the primordial symmetry-breaking transitions that distinguished the fundamental forces and gave particles their masses by means of the Higgs mechanism. And in its most generalized form, statistical physics is promising to offer insights into phenomena once considered outside the physicist’s domain: traffic flow, economics, cell biology and allometric scaling (the relation of biological functions to body mass), to name a few.

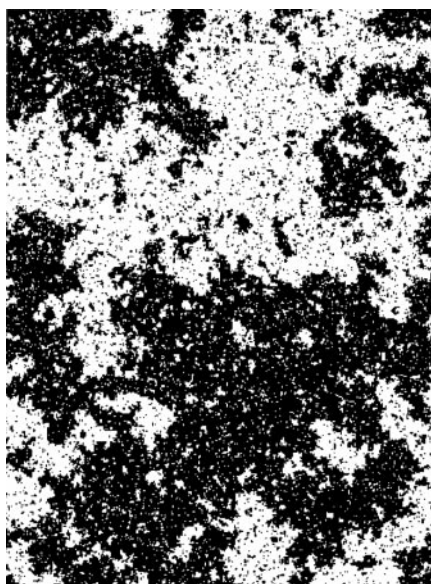


Figure 1 The Ising model at the critical point. Each site on this two-dimensional lattice can adopt one of two states — black or white, corresponding to ‘up’ or ‘down’ spins in a ferromagnet. At the critical point, neither state predominates, and fluctuations occur on all length scales. (Courtesy of Alistair Bruce, University of Edinburgh.)

Critical ideas

‘Phase transition’ is today a debased term — like the classical equivalent of ‘quantum leap’, it tends to attach itself to any abrupt change in a system’s behaviour. Does a single molecule, such as a protein, undergo a phase transition if it abruptly changes conformation? In the strict sense, no. A genuine transition requires that there be some singularity in a thermodynamic potential (such as the Gibbs free energy), which in itself requires that one can characterize the states of the system in a ‘thermodynamic limit’ of infinite system size. But too much generality may be no bad thing, if it drives home the message that phase transitions occur not only when a liquid freezes or evaporates but also throughout the (once sub-microscopic) Universe as it cools, or in a superfluid as its viscosity vanishes. The point is that phase transitions are global and abrupt — they show matter behaving at its most nonlinear, with effects quite out of proportion to cause.

That such a versatile discipline as statistical physics should have remained so well hidden that only aficionados recognize its importance is a puzzle for science historians to ponder. (The topic has, for example, in one way or another furnished 16 Nobel prizes in physics and chemistry.) Perhaps it says something about the discipline’s humble beginnings, stemming from the work of Rudolf Clausius, James Clerk Maxwell and Ludwig Boltzmann on the kinetic theory of gases. In attempting to derive the gas laws of Robert Boyle and Joseph Louis Gay-Lussac from an analysis of the energy and motion of individual particles, Clausius was putting thermodynamics on a microscopic basis. But from a modern perspective, his programme was deeper still: he was attempting to understand the collective behaviour of interacting, many-body systems. This, it might be argued, is the defining objective of statistical physics in all its guises.

At least with (dilute) gases one can afford to neglect interparticle attractive forces with some justification. Phase transitions enter into the picture, however, when those forces are included. Johannes Diderik van der Waals, who introduced such forces in a heuristic manner using what we would now call a mean-field theory, found that he could describe the gas–liquid transition. In van der Waals’ theory, the particles have a hard repulsive core and an infinitesimally small attraction of infinite range (although this is not the way the Dutchman expressed it).

Van der Waals was awarded the Nobel prize in 1910 and is regarded as something of a founding father for statistical physics. So far did his vision penetrate that in 1998 the physicist Ben Widom, in his Boltzmann Medal address, could still ask “what do we know that van der Waals did not know?”, and answer “not very much”¹. In particular, he was well aware not only of the gas–liquid critical point (which his equation predicts) but also of the existence of critical exponents, which describe mathematically how various properties vanish or diverge at the critical point. It is at this unique point in the ‘phase space’ of temperature, pressure and density that a liquid and gas cease to be distinct and separated by a phase transition: above the critical temperature, there is only one fluid

phase. Thermally driven fluctuations in density of the liquid and gas (caused by the mere randomness of particle motions) become increasingly pronounced as the critical point is approached; and their range becomes infinite exactly at criticality, dragging with them so-called 'response functions' such as the fluid's compressibility.

From the 1960s to the 1980s, nothing obsessed statistical physicists more than critical points. It seems strange, at first glance, that so much attention should be focused on a specific location in the phase diagram; but the reasons are twofold. First, the behaviour of a system at its critical point also determines its behaviour in the broad vicinity too, within the so-called critical region. The fluctuations that overwhelm the system at the critical point remain significant well beyond it; one of the reasons why the (controversial) idea of a high-pressure, low-temperature liquid-liquid critical point in water² is so stimulating is that it might be expected to affect the liquid's behaviour under everyday conditions.

But second, behaviour of a system at a critical point is like a badge of identification: it reveals kinships between different systems. Liquid-gas criticality and the behaviour of some magnets at their Curie point (the temperature above which they lose their ferromagnetism) have numerically equal critical exponents, and both can be modelled by the so-called Ising model, a lattice of two-state spins. Commonality of critical exponents gives rise to the idea of universality — that is, there are generic models in statistical physics that describe a variety of apparently different many-body systems. This means that solving one statistical mechanical problem generally delivers solutions for several others at the same time; it also implies that, fundamentally, many-body behaviour is determined only by broad-brush features such as the range of interparticle forces, the dimensionality, and the nature of the 'order parameter' whose abrupt change from zero to a non-zero value defines the transition.

Actually obtaining numerical values for critical exponents from first-principles theory is, however, another matter. Mean-field models offer analytic solutions, but they are strictly approximations in anything less than four spatial dimensions. Lars Onsager's *tour de force* in 1944 was the exact solution of the two-dimensional Ising model, providing exact numbers for the critical exponents. But the three-dimensional Ising model continues to rebuff the advances of theoreticians, and may be analytically insoluble.

On the other hand, Kenneth Wilson's renormalization group theory provided in the 1960s and early 1970s a methodology for computing critical exponents numerically³. It also pointed to the scale-invariant behaviour of fluctuations at the critical point: the fact that they occur on all length

scales (Fig. 1). (Gaussian random noise, in contrast, generates fluctuations of a characteristic average amplitude.)

The principle of renormalization is to capture the fundamental probability distribution of the different states of the system by 'coarse-graining' — a kind of mathematical squinting that eliminates extraneous detail. In a lattice model such as the Ising model, where the particles occupy sites on a regular grid, this involves calculating the average state of blocks of sites of specified size. Thus, whereas in the Ising model each site is assigned a two-state variable ('up' or 'down' spin, say, represented by the black and white squares in Fig. 1), the renormalized system contains a broader spectrum of averaged 'block variables'. The interaction strength between blocks is rescaled accordingly.

Progressive rescaling at different block sizes smoothes out ever-larger fluctuations. At temperatures either side of the critical temperature, this makes the system 'look' ever further from criticality — it begins to resolve itself into one equilibrium state or the other. Exactly at the critical point, however, rescaling creates a patchwork rather like that in Fig. 1 (but with grey squares too) no matter how large the blocks become. The probability distribution of block variables settles down to an invariant form, peaked on 'black' and 'white'. The way in which this 'configuration flow' evolves with changing length scale allows one to determine the precise value of the critical exponents — which are generally different, in one, two and three dimensions, from their mean-field values.

Quantum transitions

One of the richest veins of statistical mechanics presently being mined is found at its intersection with quantum mechanics. In particular, the many-body behaviour of electrons in condensed matter is extraordinarily rich. Correlated behaviour of electrons, in which they display a degree of collective or coherent dynamics, produces for example superconductivity, the integer and fractional quantum Hall effect (quantization of the Hall resistance, a measure of the voltage generated transverse to a current in a flat conducting sheet by an applied magnetic field), heavy-fermion behaviour (where conduction electrons acquire a very large 'effective mass'), spin density waves and colossal magnetoresistance (CMR: a strong variation in electrical resistance owing to an applied magnetic field). All of these collective phenomena have in recent years been shown to underlie unexpected and potentially useful properties of novel materials: CMR, for instance, could potentially furnish highly sensitive read-out heads for magnetic memories.

Superconductivity and superfluidity were

always very evidently phase transitions — abrupt changes from a resistive to a non-resistive state, from a viscous to a non-viscous fluid. It was not until 1938, however, that the connection to statistical physics was made, when Fritz London pointed out⁴ that superfluidity in liquid helium might be the result of a kind of quantum condensation transition, in which the particles of the system become bosonic (that is, having integer spin) and so capable of occupying a single quantum state. It became generally accepted that these phenomena were examples of Bose-Einstein condensation, although the exact connection remains murky.

In superconductivity the fermionic (spin-1/2) electrons become bosons by forming Cooper pairs, a many-body effect that (in the conventional low-temperature superconductors) results from an effective attraction mediated by the electrons' interaction with lattice vibrations (phonons) in the crystal. The full details of that process were determined in the 1950s by John Bardeen, Leon Cooper and Bob Schrieffer⁵. It is significant that this is one of the many phase transitions that can be described in an approximate way by the phenomenological theory developed by Lev Landau and Vitaly Ginzburg in the 1950s. This stemmed from a very general mean-field model of phase transitions proposed earlier by Landau, whose contribution to this area of condensed-matter physics was pivotal. Landau's theory — a kind of generalized and augmented all-purpose van der Waals equation — remains the first port of call for any simplified model of a phase transition.

The observation of Bose-Einstein condensation in atomic matter, seen for the first time in cooled sodium atoms in 1995⁶, was itself another instance of a quantum phase transition — made possible now by laser cooling techniques.

But is there any systematic way to accommodate quantum effects into the existing framework of statistical mechanics? The intense interest in correlated electron systems has now provided something of the kind. Conventionally, phase transitions are induced by changes in temperature — ice melts when heated, hot iron orders magnetically when cooled. All this is driven by changes in thermal fluctuations. 'Pure' quantum phase transitions, meanwhile, take place at zero temperature, and are induced by altering some other parameter, such as the strength of an applied magnetic field, that affects quantum fluctuations. Classical fluctuations are frozen out at zero kelvin, but quantum fluctuations, which are a consequence of the uncertainty principle, remain. Their effect can be enhanced by altering some variable that alters the particles' (generally the electrons') state of localization, which is the equivalent in this case of altering the temperature.

Quantum phase transitions provide a wider screen on which to scrutinize the still-enigmatic high-temperature superconductivity of layered copper oxide compounds. Superconductivity is now recognized as just one of the manifestations of the rich physics of the correlated electrons in these systems. There are magnetic interactions between the copper ions, which bear spins by virtue of their unpaired electrons. The prototype, lanthanum strontium copper oxide (close to the material studied by Georg Bednorz and Alex Müller in 1986⁷), is now recognized as an example of a diluted quantum antiferromagnet: diluted, that is, by the 'dopant' strontium, which contributes positively charged mobile charge carriers (holes) to the copper-oxide layers. These holes may segregate into stripes, and the intervening phase, an insulating antiferromagnet, undergoes a magnetic quantum phase transition at some critical doping level.

The behaviour of such systems at energy scales corresponding to temperatures above 100 K or so is now well understood; but the transition to a superconducting state at lower temperatures involves additional interactions of the electronic and spin degrees of freedom that have yet to be elucidated. At least one Nobel prize lies in wait for this enterprise.

Beyond equilibrium

Some of the major unresolved questions about the behaviour of both everyday and exotic matter are not so much a matter of being as of becoming. How does change in these many-particle ensembles occur? Traditionally, thermodynamics has concerned itself with equilibrium states. The particles might be in frantic motion, but the macroscopic properties are constant observables of the states that minimize free energy: pressure, density, temperature, magnetization and so forth. But how do systems actually get from state to state (particularly in cases of violent change, such as explosions or fracture)? And what about systems that do not reach equilibrium at all?

First, there is the issue of irreversibility, which is to say, of the Second Law of Thermodynamics. Thermodynamic change has a directionality to it: entropy increases. The question is why. The answer most would give is Boltzmann's, which is purely probabilistic: entropy being related to the number of configurations available to the ensemble, those states will be achieved that have overwhelmingly greater probability. In the words of US physicist Josiah Willard Gibbs, "the impossibility of an uncompensated decrease in entropy seems to be reduced to an improbability".

But questions, even lingering doubts, persist. At the microscopic scale, the interaction between two particles is generally

considered to be time-reversible: an encounter can be run in reverse without becoming a physical nonsense. So in what manner does the Second Law break down in going from irreversible macroscopic systems to reversible microscopic ones? And what is the relationship to the chaotic dynamics evident even in few-particle (sometimes even two-particle) systems? Perhaps instead there is some feature of the time-dependent evolution of probability distributions that forces irreversibility irrespective of scale? In other words, might irreversibility be built in at the microscopic level (contrary to the normally accepted tenet of microscopic reversibility)? Furthermore, the Boltzmann position requires that, as Richard Feynman put it, "For some reason, the universe at one time had a very low entropy". But for what reason?

Second, there is the question of whether a thermodynamic (and corresponding microscopic) description of non-equilibrium systems can be developed that mirrors the mature description of equilibrium states. As most processes of interest, from atmospheric circulation to cell metabolism, take place out of equilibrium, the issue is a pressing one. A key question is whether variational or minimization principles exist for selecting the most stable dynamic state, as is the case at equilibrium. Several have been suggested, amongst them Ilya Prigogine's minimal rate of entropy production for systems close to equilibrium; but there is no consensus, and at least some indication (the late Rolf Landauer's 'blowtorch theorem'⁸) that there can in fact be no universal principle of this sort, independent of a system's past history, away from equilibrium. In short, it matters not only what state a system is in, but how it got there.

In fact, for all that one can postulate that entropy production must be positive in non-equilibrium processes (so that the Second Law is not violated if and when equilibrium is reached), it is not even clear how to calculate the production rate because there is no generally accepted definition of entropy away from equilibrium. Boltzmann's $S = k \ln W$ (relating entropy S to available microstates W) will not give it to us, nor will any other general law. While that is so, a quantitative non-equilibrium statistical mechanics remains hard to formulate. One possible way forward is to use as the appropriate variables of dynamical steady states the time-invariant probability measures developed in the 1970s by Yakov Sinai, David Ruelle and Rufus Bowen.

But as yet, the theory of non-equilibrium remains largely a heuristic one. It is evident that lack of equilibrium does not imply lack of structure — many if not most of the richest patterns in nature (Fig. 2) are formed out of equilibrium. Microscopic models such as reaction-diffusion schemes do a fair

job of capturing some of this behaviour, and 'toy' equations such as the Swift-Hohenberg equation provide a general description of some of the symmetry-breaking processes that occur⁹; but non-equilibrium systems seem particularly prone to the influences of boundary conditions, defects and noise, and remain resistant to too much generalization.

All the same, some useful broad principles have appeared, among them the idea that certain non-equilibrium systems have a kind of imposed criticality said to be self-organized, in the sense that they will constantly return to a precarious critical state following some transient instability such as a landslide¹⁰. The characteristic statistics of such systems, dubbed $1/f$ behaviour because of the inverse relationship between the size and frequency (f) of a fluctuation, provides a kind of fingerprint that alerts the observer to the likely appearance of scale-invariant structure and dynamics. Self-organized criticality seems to be a promising candidate for a general mechanism capable of forming the fractal structures so prominent in nature, from mountain ranges to the large-scale structure of the Universe. There is a clear kinship with the fractal 'optimal channel networks' that have been posited¹¹ as models of real river drainage networks. These are formed under the assumption that the evolution must minimize the rate of potential-energy dissipation in the water flow. The resulting topology of the drainage basin shows non-random $1/f$ statistics. To what extent this model captures the essential features of real river systems is still a matter of debate.

And it remains to be seen whether self-organized criticality itself will prove to be generalizable to forest fires, epidemics, solar flares and the countless other physical systems in which such statistics have been reported. But the canonical example, the sand pile, points to another growth area in studies of the behaviour of matter — granular media. Capable of both solid-like and fluid-like behaviour, dominated by dissipative collisions and essentially independent of temperature, granular media represent a new class of problem for many-body theorists. There is still no general physical framework, comparable to the kinetic model of gases, able to predict the stupefying range of collective and self-organized behaviour exhibited by grains in motion: size segregation in vertical shaking and landslides, formation of ordered patterns¹² (Fig. 2), liquefaction, hysteretic angles of instability and so forth. Quite aside from the relevance of much of this for industrial powder processing, there is the matter of whether it might cast light on the geomorphology of grainy media distributed by wind and wave.

Traffic flow is now understood to be a special case of granular flow, and exhibits dynamic states that bear striking analogy to

the equilibrium states of matter. In low-density 'free' flow, each vehicle moves more or less independently, like a gas. A traffic jam at high densities is resolutely solid-like, with equidistant particles trapped in near or total immobility. But jams rarely nucleate from free flow; instead, it seems that a dense, congested yet mobile state intervenes, the highway equivalent of a liquid state. Changes from one state to another seem to have the abrupt character of a first-order phase transition (like freezing or melting), relying on the presence of fluctuations to nucleate the change.

Perhaps the hardest of non-equilibrium problems — turbulence — remains as obstinate as ever. The mathematician Sir Horace Lamb's famous quip — that he was more optimistic about receiving heavenly enlightenment on quantum electrodynamics than on turbulence — proved prophetic in terms of which would yield first, and turbulence still retains something of its reputation as a 'graveyard of theories'. From a statistical physical perspective, the situation is nightmarish: every parcel of fluid acts non-trivially on the others, there is structure at all length scales, dissipation is very strong, and the solutions are wholly time-dependent. Ideas from critical phenomena might prove helpful for characterizing the scaling behaviour of turbulent flows. Yet questions remain, for example, about just how many distinct regimes of thermal turbulence (as a function of Rayleigh number, proportional to the temperature gradient) there are. Each regime is characterized by a specific scaling law relating Rayleigh number to heat transport; recent predictions of an 'ultrahard' regime relevant to atmospheric dynamics have been met with conflicting experimental results¹³.

Even in equilibrium, a strong element of disorder in a system complicates the picture. The glass transition withholds some of its mysteries still, exemplifying a whole range of systems — among them spin glasses and folded proteins — in which the problem is that of finding a global energetic minimum in a rugged 'energy landscape'. How this landscape is explored as a liquid is lowered towards its glass transition is in need of further clarification before a truly thermodynamic description of this transition can be developed. The conceptual tools needed for the job are also being usefully brought to bear on the vexed issue of protein folding. This compares with a glass in the sense that the system of particles (in this case interacting residues on the polypeptide chain) has a great many configurations that correspond to local free-energy minima, but only one — the native fold — that equates with the global minimum. The protein experiences 'frustration' as it folds: an amino-acid residue cannot simultaneously optimize interactions with all its neighbours. Mapping this situation onto the case of frustrated

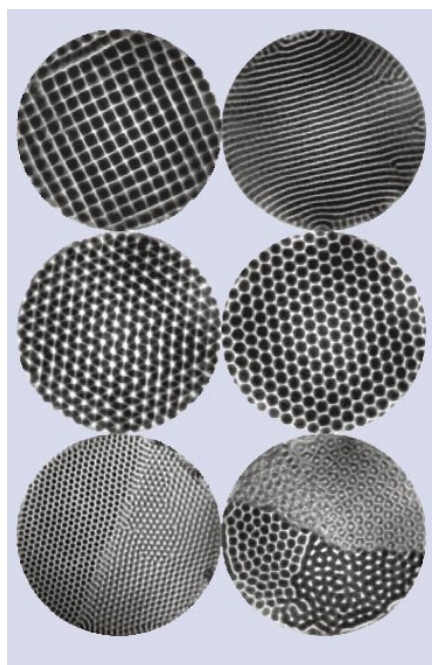


Figure 2 Complex, ordered patterns form in vertically oscillated thin layers of grains. (From ref. 12.)

magnetic spin glasses known in condensed-matter physics has helped to explain the wide range of protein folding speeds (a function of the degree of frustration) and has provided the useful concept of a folding 'funnel', the broad basin in the energy landscape that surrounds the deep well of the native fold¹⁴.

Is physics rigorous?

It can be remarkably hard to prove rigorously the development of genuine long-ranged (crystal-like or ferromagnetic) order in an equilibrium phase transition — particularly if the degrees of freedom are continuous rather than discrete (for example, in Heisenberg rather than Ising models of magnetic states, where the spins can take any orientation). We 'know' from numerical modelling that such systems do adopt long-range order in appropriate circumstances; but we cannot prove that they do so rigorously. Physicist Elliott Lieb's comment on this issue applies equally to many areas of physics: "One might ask why one should bother to prove rigorously what is 'physically obvious' and the answer is that 'Not everything that is obvious is true and the ability to prove something interesting about a model usually requires an additional degree of physical understanding that goes beyond intuition; in other words, we learn something new about physics.'"¹⁵

Such rigour is, however, surely an impossible dream when one comes to apply the elegance of statistical physics to the orchestrated chaos we call life. Despite the proven value to cell biology of some concepts from the study of phase transitions — such as the

entropic effect of fluctuations on interactions of lipid membranes — there remains much scepticism as to whether any biological phenomena can arise from the sort of collective, emergent behaviour of statistical, interacting ensembles rather than the closely controlled protein relays to which cell biologists are accustomed. Yet statistical physics must inevitably provide the baseline even in the cell: proteins may phase-separate and membranes may adopt equilibrium conformations unless actively opposed. The recent interest in thermal ratchets¹⁶ — Brownian systems that achieve directional motion by virtue of operating in an asymmetric underlying potential — attests to the contributions that microscopic physical models might make to cell biology. These models may or may not, in the end, have much to do with the way that motor proteins work; but they demonstrate that physics offers creative solutions to adaptive biological systems. The same can be said for the idea that stochastic resonance — the (counterintuitive) noise-induced amplification of a signal¹⁷ — is exploited in biological signal transduction. There now seems to be good evidence that this mechanism has some role in neural processing, and one might even be surprised if it did not turn out to be more general. In both of these cases, we see how noise — inevitable in any environment — can, with the right adaptation, serve a functional purpose.

The considerable redundancy evident in the cell's machinery — so that cells continue happily with disabled genes that were supposedly central to their survival — suggests the need for some kind of nonlinear collective modelling of the interactions among its components. This is the kind of thing that physicists have been doing for years, and in increasingly complex systems. Cells provide perhaps the ultimate challenge, although the modelling will have to be dosed with a strong element of biological good sense.

Philip Ball is a consultant editor for *Nature*.
e-mail: p.ball@nature.com

1. Widom, B. *Physica A* **263**, 500 (1999).
2. Mishima, O. & Stanley, H. E. *Nature* **396**, 329 (1998).
3. Bruce, A. & Wallace, D. in *The New Physics* (ed. Davies, P.) (Cambridge Univ. Press, 1989).
4. London, F. *Nature* **141**, 643 (1938).
5. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Phys. Rev.* **108**, 1175 (1957).
6. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. *Science* **269**, 198 (1995).
7. Bednorz, J. G. & Müller, K. A. *Z. Phys. B - Cond. Matt* **64**, 189-93 (1986).
8. Landauer, R. *Physica Rev. A* **12**, 636 (1975).
9. Cross, M. C. & Hohenberg, P. *Rev. Mod. Phys.* **65**, 851 (1993).
10. Bak, P., Tang, C. & Wiesenfeld, K. *Phys. Rev. Lett.* **59**, 381 (1987).
11. Rodriguez-Iturbe, I. & Rinaldo, A. *Fractal River Basins* (Cambridge Univ. Press, 1997).
12. Melo, F., Umbanhowar, P. B. & Swinney, H. L. *Phys. Rev. Lett.* **75**, 3838 (1995).
13. Glazier, J. A., Segawa, T., Naert, A. & Sano, M. *Nature* **398**, 307 (1999).
14. Wolynes, P. G. & Eaton, W. A. *Physics World* **12**(9), 39 (1999).
15. Lieb, E. *Physica A* **263**, 491 (1999).
16. Astumian, R. D. *Science* **276**, 917 (1997).
17. Wiesenfeld, K. & Moss, F. *Nature* **373**, 33 (1995).

Keeping time

How finely can time be measured? The answer might define our capacity to conduct precise tests of general relativity, as demonstrated already in the observed increase in clock rate with distance from the Earth's surface, or the measurement of relativistic effects in binary pulsar systems. Spacecraft navigation, interferometric radio astronomy and Global Positioning System geodesy are all dependent on timekeeping that taxes current technology. And accurate measurements of the second define other standards, such as the units of distance and voltage.

Measuring time is about monitoring oscillations, whether of the pendulum, the quartz crystal or the caesium atom. By relying on atoms for our frequency (and hence time) standards, we may at least avoid placing ourselves at the mercy of the machinist who made the device; but on the other hand, we fall foul of the Uncertainty Principle. Atomic transitions, of which the hyperfine splitting of energy levels in caesium-133 furnishes us with today's standard second, happen at a frequency that can ultimately be well defined only to the degree that the excited state is long-lived. The shorter the lifetime — or the duration of the measurement — the greater the uncertainty in energy (and frequency) of the transition.

So to get a more accurate result, one needs a more leisurely attitude to measurement. This is the principle behind the 'atom fountain', a device that first cools atoms within an optical 'molasses' of three crossed laser beams before gently launching them vertically so that they fall back under gravity. The atoms can then be probed outside the perturbing influence of the laser field.

Towards the top of the fountain, the atoms pass twice through a microwave cavity: once on the way up, once on the way down, separated by about a second. Invented by Norman Ramsey in 1949, this 'separated oscillatory-field method' is applied to a linear atomic beam in current atomic clocks. But the atom fountain permits a much longer span between pulses, and so a greater narrowing of the transition. In principle these devices offer a frequency determination accurate to one part in 10^{16} , which is equivalent to a variation of about one second in a billion years.

Philip Ball

Circadian clocks

How do organisms respond to their most fundamental timescale — the Earth's rotation around its own axis? With a suite of interconnected hormonal and physiological loops that make up the 'circadian timing system' or CTS.

Every CTS — whatever the organism — is made up of light receptors, an endogenous 'biological' clock and a so-called 'entrainment' system to couple one with the other. For although every cell of every biological or circadian clock behaves cyclically, these intrinsic periods are not necessarily exactly 24 hours long. The clock has to be set.

Much of the pathology and endocrinology of the circadian clock has been elucidated in the past century. In humans, for instance, many sections of the route from retina to brain to heart-muscle action, metabolism, body temperature and the organization of sleep-wake cycles have been charted. But there are still major areas of uncertainty, **especially in the newest branch of circadian biology: the identification and study of genes necessary for normal timekeeping.**

Clock genes seem to be present in all cells, even though only a subset is capable of acting as self-sustained oscillators. So do these genes have other, global, non-circadian functions in cell biology? As Russell Foster of Imperial College, London, puts it: "cellular assays have shown the way forward, now we need good functional analysis using transgenic approaches and behavioural studies".

Moreover, only a few animal systems have been studied so far in any detail — the mouse, the fly and the fungus, *Neurospora*. This has revealed some remarkable similarities but also some exciting differences. Comparisons of additional species will surely tell us more about the evolution and adaptive significance of circadian systems and their interactions with their environments. All of which will hopefully feed back into a fine-tuned understanding of, and ability to treat or even to manipulate, the one timepiece that we all care about — the human body.

Sara Abdulla

The challenge of conservation

Some disciplines move faster than others, but if there is one moving much faster than its practitioners would like, it is conservation biology. In the next century, conservationists will be less interested in pristine wilderness — for such will hardly exist — but in the success with which human beings manage the biosphere to the benefit of all its inhabitants, including humans. The world is currently in an unprecedented situation in which *Homo sapiens* — just one species out of millions — sequesters 45 per cent of the planet's net terrestrial productivity and 55 per cent of its fresh water. Effective biodiversity systems management will become an imperative if only to maintain the quality of human life, let alone that of other species.

In the coming decades, conservation biology will become a close analysis of the art of compromise, as humanity seeks to understand the global ecosystem as a participant, not as an observer. "Within a century," says Peter Raven, director of the Missouri Botanic Garden, "the world will be a patchwork with varying degrees of biological richness, and the way people respond locally to challenges will determine the nature of that compromise."

What will we have lost by 2100? Raven estimates that two-thirds of all species will disappear if current trends continue, but is nevertheless optimistic. "We can do better," he says. **"No major category of organisms, or local community of organisms, need disappear completely."**

Ecologists are acutely aware that the maintenance of ecosystem function transcends the conservation of any particular species, so high-technology initiatives such as 'ex situ' conservation — maintaining banks of germplasm outside the context of a habitat — may become a side issue.

"High technology is unlikely to solve most of the major global environmental issues that we face,"

says David Tilman of the University of Minnesota. "Rather, lifestyle changes, most likely imposed by legislation, will be needed. Such legislation will be socially sustainable only once citizens understand their long-term dependence on biodiversity and ecosystem services." And Raven adds that it is "better for now to follow Aldo Leopold's dictum: 'the first rule of intelligent tinkering is to save all the cogs and wheels'".

Henry Gee



Physics at the Planck time

Superstring theory, perhaps the most notorious attempt to unify general relativity with quantum mechanics, was once described as a piece of twenty-first-century physics that fell into the twentieth century. Certainly, physicists struggling to marry these reluctant partners have long sensed that the crucial piece of the puzzle has yet to be invented. But a theory of quantum gravity remains more a question of tidiness than anything else — quantum effects on space-time become important only on time and length scales well out of experimental reach. A quantum field theory of electromagnetism has plenty of practical relevance for applied physics, and such a field theory for the nuclear forces generates no shortage of testable predictions. But gravity yields to the quantum rod only at the so-called Planck scale: over distances of about 10^{-33} cm and times of 10^{-43} s.

A physics to match the Planck timescale is the biggest challenge to physicists in the coming century. **This is the timescale relevant to the graviton, the putative carrier of the gravitational force.** And naturally enough, only such a theory can extend our understanding of the Big Bang inside the first 10^{-43} seconds of existence.

What will a theory of quantum gravity do to space-time? Will it be like John Wheeler's 'quantum foam', furiously alive with ephemeral black holes and worm holes? Or will it be more like Abhay Ashtekar's fabric of woven loops? Will particles become strings, and will the strings be supersymmetric? How many extra dimensions will turn up, curled out of sight over the Planck distance?

And will we ever be able to test such a theory? One quantum-gravity researcher has confessed that "it seems highly unlikely that a machine will ever be built with which these minute distances can be studied directly". For guidance, it seems that physicists may have to fall back on the old stand-bys: elegance and beauty. **Philip Ball**

The speed of computers

One version of Moore's Law, predicted by Intel co-founder Gordon Moore in 1965, has it that computer chips double in speed every 18 months. Nothing moves any faster, of course — the chips simply become more powerful by packing more devices into the same space. Speed is about miniaturization. **Recent research on the ultimate size limits for silicon technology have fuelled speculation about whether Moore's Law is headed for a crash.** By 2012, at the current rate of shrinkage, transistors will have become so small that the silicon oxide films will no longer guarantee adequate insulation between conducting regions: they become switches that cannot be fully turned off. What happens then to our expectation that the next generation of machines will be better, meaner and faster?

All manner of factors determine how quickly, say, the *Nature* web page downloads onto your screen, many of them scarcely connected to the component density on the chips inside the processor. In some parts of the world it can take ten minutes or more. Yet you would be waiting a lot longer if the data were streaming down old copper telephone cables rather than glass optical fibres. Fibre-optic networks have a greater carrying capacity that speeds up transmission by one or two orders of magnitude relative to electrical lines. But that is only a fraction of what should be possible with photonic technology.

And the electronics at either end are soon to be stretched to their limit to cope with the transmission speeds that photonics offers. Gigabit-per-second distributed networks have already been demonstrated, and laser diodes can in principle blink out 100 Gbits per second; but electronics lose the ability to cope at half that rate. **All-optical networks will then be required, and networks working at 100 Gbits per second are already in development.** At some point, the light will need to go on a chip, in photonic integrated circuits.

But photonics is an embryonic technology, and plagued with obstacles in comparison with mature electronics. No clear rival to the transistor has yet emerged from this or any other direction, such as single-electron devices and molecular electronics. So perhaps we will have to adjust our expectations as the information revolution reaches a plateau. Moore predicts that we have about two generations of computers left with current technologies, and "beyond that, life gets very interesting". **Philip Ball**

Lifespan extension

Eat more fibre! Consume less cholesterol! Drizzle more olive oil! Have fewer children later in life! Drink a glass of red wine a day! And so it goes on. Rarely a day goes by without the emergence of a new 'theory' in that most seductive realm of scientific advances — lifespan extension.

Chromosome structure hints that our cells have genetic 'fuses' that mete out their allotted time. Antibiotics, vaccinations and antiseptics have brought about hikes in life expectancy in the developed world which have raised the average lifespan to almost double that at the last *fin de siècle* — nigh-on 80 years and still rising — as this century draws to a close. The intractability of several cancers such as Hodgkin's lymphoma and leukaemia has been tamed and many others look to follow soon; the hard nut of cognitive decline is on the verge of being cracked with drugs and neuron replacement; and workable synthetic organs and xenotransplants are around the corner.

But the current obsession with all things genetic, molecular, high-tech and expensive risks missing the point that both the World Bank and World Health Organization (WHO) make patently clear. That is, **for the majority of the planet the most powerful lifespan-extending technologies of all — good, basic public health and preventative medicine — are still a far-off dream, even if they are, scientifically speaking, old hat.**

As we usher in the third millennium, the WHO estimates that the citizens of the world's poorest countries (who make up two-fifths of the global population) have almost half the life expectancy — at under 50 years — of those living in the richest nations. And that gap is widening, not narrowing,

often in both directions. One third of the planet's population do not have secure access to safe water or sufficient food, let alone drugs, immunizations, or even the simplest health education. Elsewhere, billions of dollars are being spent teasing out the intricacies of gene therapy and memory enhancement. Addressing this inequity is surely one of the greatest and most pressing challenges facing us all, if not as academics, then as people. **Sara Abdulla**

Flashes in femtoseconds

For the molecule, time ticks in picoseconds, or even femtoseconds — respectively 10^{-12} and 10^{-15} s. On such timescales, a molecule tumbles through space, an atom vibrates back and forth in a crystal, a chemical bond is forged or broken. Yet the development of pulsed lasers blinking in a femtosecond staccato of flashes has enabled these fundamental molecular processes to be charted, revealing the rise and decline of ephemeral intermediates under the furiously strobed illumination.

Femtosecond spectroscopy has shown us the dynamics of bond breaking — but what of the structural aspects? When will we be able to map out the coordinates of individual atoms, as, for example, a protein docks onto the cell surface or as a chlorine radical eliminates an ozone molecule? When will we see the first atomically resolved diffraction pattern of a true transition state, and watch frame by frame as vibrations carry atoms into new unions?

This demands a marriage of the techniques of ultrafast spectroscopy with those of X-ray diffraction. The probe beam becomes a pulsed X-ray source, its flashes brief enough to 'freeze' the atomic motions yet bright enough to provide a discernible diffraction pattern. Current synchrotron sources can be pulsed at around 50 picoseconds, whereas ultrashort laser pulses can excite bursts of X-rays from suitable targets that could, in principle, be over within 100 femtoseconds.

These pulses must be synchronized with an optical pumping beam, which gets the reaction underway. An alternative is to diffract electrons rather than X-rays. Diffraction on the 'molecular' timescale of femtoseconds is an infant discipline which promises wonders once perfected, but which is capable right now of only the crudest of impressionistic sketches: blurred images of lattice dynamics, showing evidence of rapid change but without a single molecule (let alone an atom) in focus. The static photography of the Braggs has yet to produce its first movie. **Philip Ball**

Arrhythmias

'His heart raced'. 'Her heart skipped a beat'. 'My heart beat sluggishly in my breast.' Literature is peppered with such lines. Why? Because, as many a dramatic writer has noted, a disturbance in that most basic 'rhythm of life' — the normal human heart rate of between 60 and 100 beats per minute — is a sign that something is amiss.

Normally a heartbeat starts in the right atrium of the heart, in a specialized group of pacemaker cells known as the sinus node. This node sends an electrical signal, which spreads through the atria and ventricles of the heart, following a precise route that induces the heart to beat in the most efficient way.

If the sinus node develops an abnormal rate or rhythm, if the conduction pathway of its electrical signals is faulty or if another part of the heart takes over as the pacemaker, arrhythmias — as the various types of irregular heartbeat are generically known — can ensue. Arrhythmias may be due to heart disease, ageing, medications, metabolic imbalances or other genetic or infectious medical problems. And the ineffective blood pumping that they give rise to can cause fatigue, dizziness, light-headedness, fainting spells and, ultimately, strokes and death.

Oscillations in the cells of the sinus node operate by setting up a cyclic decay of the difference between the charge on the outside of the membrane and the inside. This causes the membrane to fire spontaneously, which raises its conductance and allows electrolytes to rush in, thus reversing the potential across the membrane. This reversed charge propagates to neighbouring cells, causing the same sequence of events therein and hence a concert of muscular contraction.

But it is currently unclear exactly how cyclically expressed genes and proteins give rise to cyclic decay in membrane potential in the heart's muscle cells. So, just as in many other areas of medical biology, the challenge now is twofold. First, to expand our knowledge of the genetics and cell signalling processes that gives rise to healthy heartbeats. And second, to integrate this into a far richer picture of the workings, and failings, of the whole heart.

Sara Abdulla

Does the past have a future?

Hardly a week goes by without the discovery of some amazing fossil that will change the way we look at the history of life. **But how long can the fossil bonanza last?** After all, the Earth has only so many rocks to investigate. Will there come a time when we can say that the fossil record is substantially complete, in that the likelihood of discovering radically new forms, or significant range extensions of known forms, becomes improbable? If so, when will that time arrive? And when it does, how will this influence the practice of palaeontologists?

"I think we are more or less at that point now," says Andrew Smith of the Natural History Museum in London. "The available outcrops are pretty heavily searched within the developed nations, and major headway has been made in exploring places like Mongolia and Madagascar." His prognosis is, however, tempered by the fact that absolute completeness is an impossibility, because the vast majority of organisms do not end up as fossils: **"my belief is that we will have a substantially biased fossil record that cannot be bettered no matter where we go in the world".**

Charles Marshall of Harvard University agrees that the Earth's fossil riches have been fairly well worked out, but says that one cannot discern a point when the last fossil will have been extracted. "One can't simply extrapolate from what has been discovered because radical new finds often result from unanticipated ways of looking, often in stratigraphic intervals or rock types not typically examined. If the rate of recent discoveries, from fossil embryos to down-covered dinosaurs, are any indication there is much to be discovered yet!"

However, contrary to popular iconography, most palaeontologists do not devote most of their time to making radical discoveries. Such discoveries, says Marshall, will be — as they are now — "largely the unexpected by-products of mature research programmes designed to establish the patterns of evolution and elucidate the processes responsible for those patterns."

"The obvious priority," says Smith, "will be to establish and understand how sequence stratigraphy and sea-level change

control facies preservation, and thus control apparent biotic diversity and extinction patterns through time. Only then will we begin to get a reliable handle on how diversity has changed over time."

Henry Gee

Adapting to climate change

Legislating for the present on the basis of predictions for the future can never be other than a dicey business. The suggestion by a British member of parliament in the nineteenth century that London would be waste-high in horse excrement by the 1950s could have been seen as a call for crippling taxes on Hansom cabs, pushing cabbies out of trade while still leaving us choking on exhaust fumes. ***So should we now be legislating against those exhaust fumes, in fear that by the end of the next century the world will be a flood- and storm-prone hothouse?***

This was one of the key issues highlighted by the Kyoto convention on climate change in 1997. It is all very well to set targets for emissions reductions on the basis of climate models projected over the next century (and then in all probability to fall short of attaining them anyway), but what of the costs of fierce constraints to today's populations, particularly those who live within fragile economies?

Yet clearly we have to do something. There seems to be no serious doubt now that global warming, induced by emission of greenhouse gases through human activity, is upon us. The global average temperature has risen by around half a degree Celsius during the twentieth century, and is predicted to increase by a further 1.0–3.5 °C by 2100. Over the same periods, the seas have risen by 10–25 cm and we can expect almost a metre more at the upper limit. Storms and drought may or may not get more common, but their distribution will probably alter. Massive famine and flooding cannot be ruled out.

As it happens, the Kyoto targets may make very little impact on all of this, but the basic question remains: how can we plan for social change on timescales over which forecasts of enabling technologies become unreliable? One approach that has been advocated is to formulate policy based on assessments of human and social impacts rather than in terms of national emissions reduction targets — something that is harder to quantify but closer to people's concerns. Emissions are not the whole story; what about economic efforts to make societies vulnerable to climate change less so? And how far do political and industrial structures need re-engineering to make actual attainment of objectives realistic?

Climate modellers, meanwhile, are trying to incorporate some element of the social and economic contexts into their forecasts. They grapple with the question of whether it is better, given the uncertainties, to brake hard on emissions now, do nothing until later on the assumption that we will then be much more technologically capable of it (owing to enhanced energy efficiency and so forth), or something in between. Inevitably, such modelling is painfully contentious and strongly influenced by the assumptions that go into it. We are juggling with the unknown; and in the face of massive inertia and sharp political lobbying, that is a precarious occupation.

Philip Ball

The shape of the cosmos

Olaf Stapledon is the great unsung hero of twentieth-century science fiction. His stories are cast on scales of time and space so vast as to inspire nothing short of vertigo or terror. In *Star Maker*, his hero, in a 'hawk-flight of the imagination', journeys through the cosmos and eventually meets the Creator in his workshop. The Star Maker, it seems, is still perfecting his art. Our own cosmos is only provisional. Previous essays in Universe construction litter the room like trash. These essays come in all shapes and sizes, suggesting that we need not live in a great, expanding hypersphere, infinite in all directions.

Our conception of the shape and size of the Universe is based on timescales. The familiar 'light-year' is a unit based on the time it takes for a beam of light to get from A to B. But there is a reckoning that is purely geometric, or, rather, trigonometric. This is the 'parsec', short for 'parallax arcsecond', the distance at which the separation of the Earth and Sun (1 Astronomical Unit) would subtend an angle of one second of arc. A parsec is equivalent to just over three light-years.

But the dimensions of the cosmos are reckoned in megaparsecs — millions of parsecs. ***At the cosmic scale, the uncertainty over the rate at which the Universe is expanding means that time, distance and geometry are inextricably bound together:*** the Hubble flow is meted out in units of kilometres per second per megaparsec. On the very largest scales, time either loses all meaning or becomes subservient to geometry. Our understanding of time may depend on knowing the shape of the cosmos — the shape on the Star Maker's workbench.

In 2002, we may know. By then, we should have results from detailed maps of fluctuations in the cosmic microwave background (CMB) from two spacecraft: NASA's *Microwave Anisotropy Probe* and the European Space Agency's *Planck Surveyor*. The results will help constrain values for important cosmological parameters such as the density of matter in the Universe, and the Hubble constant that provides a scale to its expansion. Accurate knowledge of these parameters should, in turn, constrain ideas about the geometry of space — whether it is 'flat' or 'Euclidean', or if it has positive (that is, spherical) or negative (hyperbolic) curvature.

The results will also be able to tell us whether space is infinite in all directions, or is geometrically finite.

If it is finite, then some extremely weird things could fall out of the analysis. In a finite Universe, surveys of deep space could be pulling out multiple images of the same finite set of galaxies — like an observer in a hall of mirrors. One of those faraway galaxies could be our own, its light having circumnavigated the cosmos to reach us. Study of the CMB will be able to tell us whether space really is a hall of mirrors, and, if it is, something of the hall's architecture. The Universe could be an analogue of a simple sphere. On the other hand, it could have some more exotic shape, akin to a torus, or another of those shapes littering the Star Maker's garage floor.

Henry Gee

Keeping time

How finely can time be measured? The answer might define our capacity to conduct precise tests of general relativity, as demonstrated already in the observed increase in clock rate with distance from the Earth's surface, or the measurement of relativistic effects in binary pulsar systems. Spacecraft navigation, interferometric radio astronomy and Global Positioning System geodesy are all dependent on timekeeping that taxes current technology. And accurate measurements of the second define other standards, such as the units of distance and voltage.

Measuring time is about monitoring oscillations, whether of the pendulum, the quartz crystal or the caesium atom. By relying on atoms for our frequency (and hence time) standards, we may at least avoid placing ourselves at the mercy of the machinist who made the device; but on the other hand, we fall foul of the Uncertainty Principle. Atomic transitions, of which the hyperfine splitting of energy levels in caesium-133 furnishes us with today's standard second, happen at a frequency that can ultimately be well defined only to the degree that the excited state is long-lived. The shorter the lifetime — or the duration of the measurement — the greater the uncertainty in energy (and frequency) of the transition.

So to get a more accurate result, one needs a more leisurely attitude to measurement. This is the principle behind the 'atom fountain', a device that first cools atoms within an optical 'molasses' of three crossed laser beams before gently launching them vertically so that they fall back under gravity. The atoms can then be probed outside the perturbing influence of the laser field.

Towards the top of the fountain, the atoms pass twice through a microwave cavity: once on the way up, once on the way down, separated by about a second. Invented by Norman Ramsey in 1949, this 'separated oscillatory-field method' is applied to a linear atomic beam in current atomic clocks. But the atom fountain permits a much longer span between pulses, and so a greater narrowing of the transition. In principle these devices offer a frequency determination accurate to one part in 10^{16} , which is equivalent to a variation of about one second in a billion years.

Philip Ball

Circadian clocks

How do organisms respond to their most fundamental timescale — the Earth's rotation around its own axis? With a suite of interconnected hormonal and physiological loops that make up the 'circadian timing system' or CTS.

Every CTS — whatever the organism — is made up of light receptors, an endogenous 'biological' clock and a so-called 'entrainment' system to couple one with the other. For although every cell of every biological or circadian clock behaves cyclically, these intrinsic periods are not necessarily exactly 24 hours long. The clock has to be set.

Much of the pathology and endocrinology of the circadian clock has been elucidated in the past century. In humans, for instance, many sections of the route from retina to brain to heart-muscle action, metabolism, body temperature and the organization of sleep-wake cycles have been charted. But there are still major areas of uncertainty, **especially in the newest branch of circadian biology: the identification and study of genes necessary for normal timekeeping.**

Clock genes seem to be present in all cells, even though only a subset is capable of acting as self-sustained oscillators. So do these genes have other, global, non-circadian functions in cell biology? As Russell Foster of Imperial College, London, puts it: "cellular assays have shown the way forward, now we need good functional analysis using transgenic approaches and behavioural studies".

Moreover, only a few animal systems have been studied so far in any detail — the mouse, the fly and the fungus, *Neurospora*. This has revealed some remarkable similarities but also some exciting differences. Comparisons of additional species will surely tell us more about the evolution and adaptive significance of circadian systems and their interactions with their environments. All of which will hopefully feed back into a fine-tuned understanding of, and ability to treat or even to manipulate, the one timepiece that we all care about — the human body.

Sara Abdulla

The challenge of conservation

Some disciplines move faster than others, but if there is one moving much faster than its practitioners would like, it is conservation biology. In the next century, conservationists will be less interested in pristine wilderness — for such will hardly exist — but in the success with which human beings manage the biosphere to the benefit of all its inhabitants, including humans. The world is currently in an unprecedented situation in which *Homo sapiens* — just one species out of millions — sequesters 45 per cent of the planet's net terrestrial productivity and 55 per cent of its fresh water. Effective biodiversity systems management will become an imperative if only to maintain the quality of human life, let alone that of other species.

In the coming decades, conservation biology will become a close analysis of the art of compromise, as humanity seeks to understand the global ecosystem as a participant, not as an observer. "Within a century," says Peter Raven, director of the Missouri Botanic Garden, "the world will be a patchwork with varying degrees of biological richness, and the way people respond locally to challenges will determine the nature of that compromise."

What will we have lost by 2100? Raven estimates that two-thirds of all species will disappear if current trends continue, but is nevertheless optimistic. "We can do better," he says. **"No major category of organisms, or local community of organisms, need disappear completely."**

Ecologists are acutely aware that the maintenance of ecosystem function transcends the conservation of any particular species, so high-technology initiatives such as 'ex situ' conservation — maintaining banks of germplasm outside the context of a habitat — may become a side issue.

"High technology is unlikely to solve most of the major global environmental issues that we face,"

says David Tilman of the University of Minnesota. "Rather, lifestyle changes, most likely imposed by legislation, will be needed. Such legislation will be socially sustainable only once citizens understand their long-term dependence on biodiversity and ecosystem services." And Raven adds that it is "better for now to follow Aldo Leopold's dictum: 'the first rule of intelligent tinkering is to save all the cogs and wheels'".

Henry Gee



Physics at the Planck time

Superstring theory, perhaps the most notorious attempt to unify general relativity with quantum mechanics, was once described as a piece of twenty-first-century physics that fell into the twentieth century. Certainly, physicists struggling to marry these reluctant partners have long sensed that the crucial piece of the puzzle has yet to be invented. But a theory of quantum gravity remains more a question of tidiness than anything else — quantum effects on space-time become important only on time and length scales well out of experimental reach. A quantum field theory of electromagnetism has plenty of practical relevance for applied physics, and such a field theory for the nuclear forces generates no shortage of testable predictions. But gravity yields to the quantum rod only at the so-called Planck scale: over distances of about 10^{-33} cm and times of 10^{-43} s.

A physics to match the Planck timescale is the biggest challenge to physicists in the coming century. **This is the timescale relevant to the graviton, the putative carrier of the gravitational force.** And naturally enough, only such a theory can extend our understanding of the Big Bang inside the first 10^{-43} seconds of existence.

What will a theory of quantum gravity do to space-time? Will it be like John Wheeler's 'quantum foam', furiously alive with ephemeral black holes and worm holes? Or will it be more like Abhay Ashtekar's fabric of woven loops? Will particles become strings, and will the strings be supersymmetric? How many extra dimensions will turn up, curled out of sight over the Planck distance?

And will we ever be able to test such a theory? One quantum-gravity researcher has confessed that "it seems highly unlikely that a machine will ever be built with which these minute distances can be studied directly". For guidance, it seems that physicists may have to fall back on the old stand-bys: elegance and beauty. **Philip Ball**

The speed of computers

One version of Moore's Law, predicted by Intel co-founder Gordon Moore in 1965, has it that computer chips double in speed every 18 months. Nothing moves any faster, of course — the chips simply become more powerful by packing more devices into the same space. Speed is about miniaturization. **Recent research on the ultimate size limits for silicon technology have fuelled speculation about whether Moore's Law is headed for a crash.** By 2012, at the current rate of shrinkage, transistors will have become so small that the silicon oxide films will no longer guarantee adequate insulation between conducting regions: they become switches that cannot be fully turned off. What happens then to our expectation that the next generation of machines will be better, meaner and faster?

All manner of factors determine how quickly, say, the *Nature* web page downloads onto your screen, many of them scarcely connected to the component density on the chips inside the processor. In some parts of the world it can take ten minutes or more. Yet you would be waiting a lot longer if the data were streaming down old copper telephone cables rather than glass optical fibres. Fibre-optic networks have a greater carrying capacity that speeds up transmission by one or two orders of magnitude relative to electrical lines. But that is only a fraction of what should be possible with photonic technology.

And the electronics at either end are soon to be stretched to their limit to cope with the transmission speeds that photonics offers. Gigabit-per-second distributed networks have already been demonstrated, and laser diodes can in principle blink out 100 Gbits per second; but electronics lose the ability to cope at half that rate. **All-optical networks will then be required, and networks working at 100 Gbits per second are already in development.** At some point, the light will need to go on a chip, in photonic integrated circuits.

But photonics is an embryonic technology, and plagued with obstacles in comparison with mature electronics. No clear rival to the transistor has yet emerged from this or any other direction, such as single-electron devices and molecular electronics. So perhaps we will have to adjust our expectations as the information revolution reaches a plateau. Moore predicts that we have about two generations of computers left with current technologies, and "beyond that, life gets very interesting". **Philip Ball**

Lifespan extension

Eat more fibre! Consume less cholesterol! Drizzle more olive oil! Have fewer children later in life! Drink a glass of red wine a day! And so it goes on. Rarely a day goes by without the emergence of a new 'theory' in that most seductive realm of scientific advances — lifespan extension.

Chromosome structure hints that our cells have genetic 'fuses' that mete out their allotted time. Antibiotics, vaccinations and antiseptics have brought about hikes in life expectancy in the developed world which have raised the average lifespan to almost double that at the last *fin de siècle* — nigh-on 80 years and still rising — as this century draws to a close. The intractability of several cancers such as Hodgkin's lymphoma and leukaemia has been tamed and many others look to follow soon; the hard nut of cognitive decline is on the verge of being cracked with drugs and neuron replacement; and workable synthetic organs and xenotransplants are around the corner.

But the current obsession with all things genetic, molecular, high-tech and expensive risks missing the point that both the World Bank and World Health Organization (WHO) make patently clear. That is, **for the majority of the planet the most powerful lifespan-extending technologies of all — good, basic public health and preventative medicine — are still a far-off dream, even if they are, scientifically speaking, old hat.**

As we usher in the third millennium, the WHO estimates that the citizens of the world's poorest countries (who make up two-fifths of the global population) have almost half the life expectancy — at under 50 years — of those living in the richest nations. And that gap is widening, not narrowing,

often in both directions. One third of the planet's population do not have secure access to safe water or sufficient food, let alone drugs, immunizations, or even the simplest health education. Elsewhere, billions of dollars are being spent teasing out the intricacies of gene therapy and memory enhancement. Addressing this inequity is surely one of the greatest and most pressing challenges facing us all, if not as academics, then as people. **Sara Abdulla**

Flashes in femtoseconds

For the molecule, time ticks in picoseconds, or even femtoseconds — respectively 10^{-12} and 10^{-15} s. On such timescales, a molecule tumbles through space, an atom vibrates back and forth in a crystal, a chemical bond is forged or broken. Yet the development of pulsed lasers blinking in a femtosecond staccato of flashes has enabled these fundamental molecular processes to be charted, revealing the rise and decline of ephemeral intermediates under the furiously strobed illumination.

Femtosecond spectroscopy has shown us the dynamics of bond breaking — but what of the structural aspects? When will we be able to map out the coordinates of individual atoms, as, for example, a protein docks onto the cell surface or as a chlorine radical eliminates an ozone molecule? When will we see the first atomically resolved diffraction pattern of a true transition state, and watch frame by frame as vibrations carry atoms into new unions?

This demands a marriage of the techniques of ultrafast spectroscopy with those of X-ray diffraction. The probe beam becomes a pulsed X-ray source, its flashes brief enough to 'freeze' the atomic motions yet bright enough to provide a discernible diffraction pattern. Current synchrotron sources can be pulsed at around 50 picoseconds, whereas ultrashort laser pulses can excite bursts of X-rays from suitable targets that could, in principle, be over within 100 femtoseconds.

These pulses must be synchronized with an optical pumping beam, which gets the reaction underway. An alternative is to diffract electrons rather than X-rays. Diffraction on the 'molecular' timescale of femtoseconds is an infant discipline which promises wonders once perfected, but which is capable right now of only the crudest of impressionistic sketches: blurred images of lattice dynamics, showing evidence of rapid change but without a single molecule (let alone an atom) in focus. The static photography of the Braggs has yet to produce its first movie. **Philip Ball**

Arrhythmias

'His heart raced'. 'Her heart skipped a beat'. 'My heart beat sluggishly in my breast.' Literature is peppered with such lines. Why? Because, as many a dramatic writer has noted, a disturbance in that most basic 'rhythm of life' — the normal human heart rate of between 60 and 100 beats per minute — is a sign that something is amiss.

Normally a heartbeat starts in the right atrium of the heart, in a specialized group of pacemaker cells known as the sinus node. This node sends an electrical signal, which spreads through the atria and ventricles of the heart, following a precise route that induces the heart to beat in the most efficient way.

If the sinus node develops an abnormal rate or rhythm, if the conduction pathway of its electrical signals is faulty or if another part of the heart takes over as the pacemaker, arrhythmias — as the various types of irregular heartbeat are generically known — can ensue. Arrhythmias may be due to heart disease, ageing, medications, metabolic imbalances or other genetic or infectious medical problems. And the ineffective blood pumping that they give rise to can cause fatigue, dizziness, light-headedness, fainting spells and, ultimately, strokes and death.

Oscillations in the cells of the sinus node operate by setting up a cyclic decay of the difference between the charge on the outside of the membrane and the inside. This causes the membrane to fire spontaneously, which raises its conductance and allows electrolytes to rush in, thus reversing the potential across the membrane. This reversed charge propagates to neighbouring cells, causing the same sequence of events therein and hence a concert of muscular contraction.

But it is currently unclear exactly how cyclically expressed genes and proteins give rise to cyclic decay in membrane potential in the heart's muscle cells. So, just as in many other areas of medical biology, the challenge now is twofold. First, to expand our knowledge of the genetics and cell signalling processes that gives rise to healthy heartbeats. And second, to integrate this into a far richer picture of the workings, and failings, of the whole heart.

Sara Abdulla

Does the past have a future?

Hardly a week goes by without the discovery of some amazing fossil that will change the way we look at the history of life. **But how long can the fossil bonanza last?** After all, the Earth has only so many rocks to investigate. Will there come a time when we can say that the fossil record is substantially complete, in that the likelihood of discovering radically new forms, or significant range extensions of known forms, becomes improbable? If so, when will that time arrive? And when it does, how will this influence the practice of palaeontologists?

"I think we are more or less at that point now," says Andrew Smith of the Natural History Museum in London. "The available outcrops are pretty heavily searched within the developed nations, and major headway has been made in exploring places like Mongolia and Madagascar." His prognosis is, however, tempered by the fact that absolute completeness is an impossibility, because the vast majority of organisms do not end up as fossils: **"my belief is that we will have a substantially biased fossil record that cannot be bettered no matter where we go in the world".**

Charles Marshall of Harvard University agrees that the Earth's fossil riches have been fairly well worked out, but says that one cannot discern a point when the last fossil will have been extracted. "One can't simply extrapolate from what has been discovered because radical new finds often result from unanticipated ways of looking, often in stratigraphic intervals or rock types not typically examined. If the rate of recent discoveries, from fossil embryos to down-covered dinosaurs, are any indication there is much to be discovered yet!"

However, contrary to popular iconography, most palaeontologists do not devote most of their time to making radical discoveries. Such discoveries, says Marshall, will be — as they are now — "largely the unexpected by-products of mature research programmes designed to establish the patterns of evolution and elucidate the processes responsible for those patterns."

"The obvious priority," says Smith, "will be to establish and understand how sequence stratigraphy and sea-level change

control facies preservation, and thus control apparent biotic diversity and extinction patterns through time. Only then will we begin to get a reliable handle on how diversity has changed over time."

Henry Gee